A MULTIVARIATE ADAPTIVE TRIMMED LIKELIHOOD ALGORITHM

Daniel Dice Schubert

THIS THESIS IS PRESENTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT

MURDOCH UNIVERSITY

MURDOCH, WA 6150

AUSTRALIA

2005

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

Daniel Dice Schubert

Acknowledgements

I would like to sincerely thank and acknowledge the following people who helped me with this thesis.

• My supervisor Dr. Brenton Clarke for introducing me to Robust Statistics and Multivariate Data Analysis in my graduate years and for his friendship, support and encouragement.

• Will Stirling, Murdoch University IT Guru, for helping me with all my PC problems.

• Professor Ronald Butler for sending me his personal notes regarding the asymptotic theory relating to Butler et al 1993.

• Dr. Mark Lukas for his many Matlab hints to speed up my programs and his instructions with regard to LaTeX.

• Phd candidate Suzanne Brown and Dr. Martine van de Poll for their helpful ideas with my R code.

Abstract

The research reported in this thesis describes a new algorithm which can be used to robustify statistical estimates *adaptively*. The algorithm does not require any pre-specified cut-off value between inlying and outlying regions and there is no presumption of any cluster configuration. This new algorithm adapts to any particular sample and may advise the *trimming* of a certain proportion of data considered extraneous or may divulge the structure of a multi-modal data set. Its adaptive quality also allows for the confirmation that uni-modal, multivariate normal data sets are outlier free. It is also shown to behave independently of the *type* of outlier, for example, whether applied to a data set with a solitary observation located in some extreme region or to a data set composed of clusters of outlying data, this algorithm performs with a high probability of success.

Contents

List of Figures

List of Tables

Introduction

Outliers, by their very definition, pose a threat to the sensible inferences we hope to draw from the statistical appraisal of any data set. All data sets are vulnerable to outlying data, for example data may be recorded incorrectly or may consist of faulty measurements, outliers can also be freak instances of nature or evidence of multi-modality. The source and type of outlying data may be of interest, but if not expelled from the data set being analyzed, or weighted accordingly, will corrupt parameter estimates and any ensuing statistical inference.

Most outliers in the univariate setting can be exposed by a simple scatter diagram or a stem-leaf plot but can still upset estimation when one uses non-visual methods of assessment. Investigation of multivariate data sets will include samples of dimension $p > 3$ for which visual inspections are not possible. There exists Software packages specifically designed for visual analysis of multivariate data sets. Packages such as GGobi can produce powerful images from every conceivable perspective, for each combination of 3 variables in 3 dimensions. It allows the user to interact with, and manipulate, any chosen emphasis which may include the identification of extraneous points.

Outliers are an assortment of contamination. There can exist the solitary strays or scatters of stray points and even more serious are those that compose outlying clusters. Clusters with a similar shape to the majority sample data are as difficult to detect, and as dangerous to statistical inference, as concentration clusters of tiny variance compared with the main population. Samples may even be multi-modal, in a sense containing no conventional outlier, but it is imperative that the cluster configuration be exposed by an algorithm designed to locate *abnormal* data.

One has in mind a multivariate normal distribution for the population from which the data is observed. The theoretical framework for the algorithm devised for this thesis can therefore be based on an argument of Fisher Information. When one reduces the informa-

tion a data set contains, for instance by trimming observations extreme or otherwise, one necessarily increases the variance of the parameter estimates, in particular the estimate for location. It is hoped that when the data being removed from a sample is contamination data, there will be a corresponding *decrease* in the measure of the asymptotic variance for the estimate of location.

There exists many algorithms in the pursuit of the identification of outliers in multivariate data. Some algorithms focus on specifying a cut-off region whereby those observations lying beyond this cut-off are deemed outlying, while other algorithms focus on a cluster analysis of the data set in an effort to find the best grouping of data. The algorithm proposed in this thesis identifies outliers adaptively and independent of outlier type. Without the need for the 3 dimensional visual perspectives provided by GGobi, say, this algorithm analytically identifies outliers with an assessment of *all* the variables simultaneously.

Chapter 1 of the thesis discusses various prevailing robust estimation techniques for the identification of location and scale parameters. Techniques, some of which, that are used for the new proposal introduced in this thesis. It also points out the motivation for the development of robust methods as a way of cleaning samples of corrupt data, then explores a selection of existing outlier detection algorithms. Towards the end of Chapter 1, certain cluster detection methodologies are reviewed where clusters of data are viewed in the context of non-normal, outlying data.

Chapter 2 introduces the new proposal, beginning with its emanation from already existing univariate adaptive trimmed likelihood methodology. Its theoretical underpinnings based on the asymptotics of the Minimum Covariance Determinant is then canvassed along with a description of its Forward Search component. Monte Carlo simulations, involving an array of sample and outlier types, allow for the recognition that the algorithm is dependent on sample size and needs a slight modification when dealing with small samples. The simulations will also divulge the algorithms ability to identify possible multi-modality and compare further, possible modifications. Chapter 2 continues with a comparison with other existing methods, those discussed in Chapter 1 and another, already existing, adaptive

algorithm. An application of the new proposal on real data sets is then undertaken before its formal algorithmic description to finish Chapter 2.

Chapter 3 shows how this algorithm can be used to robustify both univariate and multivariate regression analysis. Other contemporary strategies are discussed before the new proposal is applied and the ensuing results compared. Chapter 4 highlights this algorithms ability to recognize multi-modal data while Chapter 5 explores its application in conjunction with Principal Components Analysis, Discriminant Analysis and Canonical Correlation Analysis.

Chapter 1

Review of Robust Estimation techniques

1.1 Statistical Distance

Suppose we let a multivariate observation of dimension p be represented by the vector random variable $\mathbf{X}^\top = (X_1, \ldots, X_p)$ and a realization of that random variable be represented by $\mathbf{x}^\top = (x_1, \ldots, x_p)$. Then a random sample of such vector observations is represented by the sequence $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of independent, vector random variables so that any particular realization of the sample is represented by $\mathbf{x}_1, \ldots, \mathbf{x}_n$. To analyze this sample of points for *outliers* it is necessary to have a measure of *distance* between them, or say, the *magnitude* of each observation relative to the others. Two fundamental parameters for such a measure are an estimate for the location of the random variable $\mathbf{X}$, using the sample mean vector $\hat{\boldsymbol{\mu}}$, and an estimate for the measure of scatter, or the scale, of the data, $\hat{\boldsymbol{\Sigma}}$, using the covariance matrix of the sample.

To be able to identify any observed sample point, $\mathbf{x}$, as an outlier we can use its distance from the mean vector, or *centroid*, with respect to the sample covariance matrix to derive the probability any observation would have such a distance with respect to the rest of the

sample.

In two or three dimensions it is always possible to plot the points to get a perspective on the shape of the data and the relative sizes and potential outlyingness of contaminant data. When analyzing higher dimensional data, graphical methods are not suitable and so we need an analytical method, such as the measure of distances discussed above, that can effectively project the necessary information onto a one-dimensional space. We could think of using Euclidean or squared straight-line distance, D say, between points so that if we consider two points, $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$, from a bivariate data set, $p = 2$, then

$$D^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2$$

and more importantly, because we need to know how far an observation is from a sample centroid $\hat{\boldsymbol{\mu}}^\top = (\hat{\mu}_1, \hat{\mu}_2)$,

$$D^2 = (x_1 - \hat{\mu}_1)^2 + (x_2 - \hat{\mu}_2)^2.$$

What is more crucial is how these distances are related to the variation, or scatter, of the data set they belong to, $\mathbf{X}^\top = (X_1, X_2)$, or an estimate of that scatter based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$. An increase in straight line distance of a variable belonging to a set exhibiting a small variation is more significant than if it was a member of a set of values with a large variation. The distances, therefore, need to be standardized, weighted inversely by a measure of the spread of the data. Such distances are referred to as *statistical distances*.

With regard to multivariate normal distributions, this measure of standardized, or statistical, distance is such that an increase in statistical distance from a mean reflects a decrease in the probability of an observation possessing such a distance, occurring. Outliers will be identified, in this thesis, as those observations *outlying* with respect to normally distributed data. If we describe the *majority* data set by a normal density then only those observations with a significantly low probability, with respect to this density, will be identified as outliers.

A multivariate normal distribution of dimension p can be defined as follows.

Beginning with a vector $\mathbf{Z}^\top = (Z_1, \ldots, Z_p)$where $Z_1, \ldots, Z_p$ are independent standard normal variables, the density of Z can be represented by the equation

$$f(\mathbf{z}) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp(-z_i^2/2) = (2\pi)^{-p/2} \exp(-\frac{\mathbf{z}^\top \mathbf{z}}{2}).$$

Now assuming we have a positive definite matrix $\mathbf{\Sigma}$ so that $\mathbf{\Sigma}^{1/2}$ is an appropriate square root of the matrix $\mathbf{\Sigma}$, then $\mathbf{X} = \boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\mathbf{Z}$ has, by the usual transformation of multivariate variables, a density given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp(-1/2(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})) \tag{1.1}$$

where $\mathbf{\Sigma}$ is the covariance matrix and $\frac{1}{|\mathbf{\Sigma}|^{1/2}}$ is the Jacobian of the transformation, which corresponds with the customary $\frac{1}{\sigma}$ for univariate normal probability densities.

The exponent in (1.1) leads to a well known example of a statistical distance, the Mahalanobis distance, devised by Mahalanobis in 1930. This is also known as Hotelling's T^2 distance, after Hotelling devised a similar statistic to that of Mahalanobis in 1931. A portrayal of such a distance when observations are randomly distributed bivariate normal, $\mathbf{X}^\top = (X_1, X_2)$ and $\text{Cov}(X_1, X_2) = 0$, and where points $\mathbf{x_1} = (x_a, x_b)$, $\mathbf{x_2} = (x_c, x_d)$ are at an *equal* Mahalanobis distance from the mean vector or centroid, $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, is given in Figure 1.1.

More generally, the Mahalanobis distance a sample vector $\mathbf{X_i}$ has from a centroid vector $\boldsymbol{\mu}$, is represented by,

$$M_i^2 = (\mathbf{X_i} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{X_i} - \boldsymbol{\mu})$$

where for p-dimensional data

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & & \sigma_p^2 \end{pmatrix}$$

In the simple case where $p = 1$ this distance reduces to

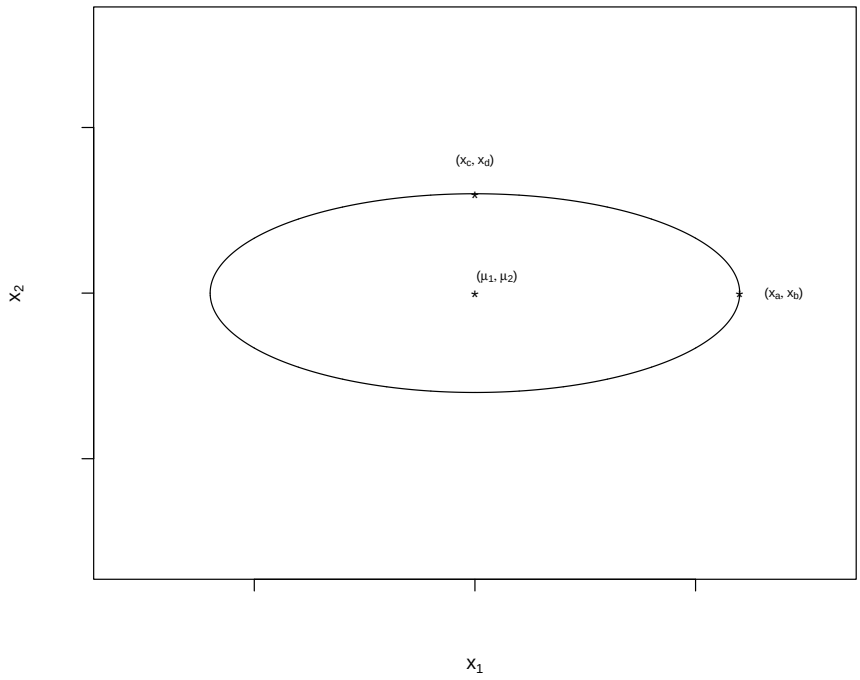


Figure 1.1: Ellipse representing an equivalent statistical distance.

$$d^2 = \frac{(X_i - \mu)^2}{\sigma^2}$$

or the squared "standardized variable".

Should $\mathbf{X}$ be bivariate normal, $p = 2$, any given contour signifying equivalent distances from a centroid can be represented by an ellipse of constant density. Indeed all multivariate normal distributions, $p > 2$, are ellipsoidal distributions whereby contours following an ellipsoidal path describe level sets of probability density functions. Regarding such distributed data it is well known that

$$M_i = (\mathbf{X_i} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X_i} - \boldsymbol{\mu}) \sim \chi^2_p,$$

the chi-squared distribution with p degrees of freedom, whence say $P(M_i^2 \leq \chi^2_{0.90,p}) = 0.90$. Consequently, for example, if $p = 2$ and $c^2 = \chi^2_{0.90,2} = 4.605$ there is a 90% chance that a random variable $\mathbf{X_i}$ will lie inside the ellipse described by the contour $M^2 = c^2$. On average 90% of the data are expected to lie inside this ellipse.

An illustration of such contours, along with a particular sample, is given in Figure 1.2. A smaller value of $c > 0$ will result in the expectation of less data being contained by the corresponding ellipse.

When identifying outliers using distance based methods we are essentially locating those observations beyond a certain statistical distance from the data centroid. We seek an outlier-region in a sense (Becker and Gather 1999) whereby the set of all observations belonging to this region are deemed suspiciously outlying. Determining the elliptical boundary that separates this region, from the region of *inlying* data, must begin with a necessarily *robust* estimate for location and scale.

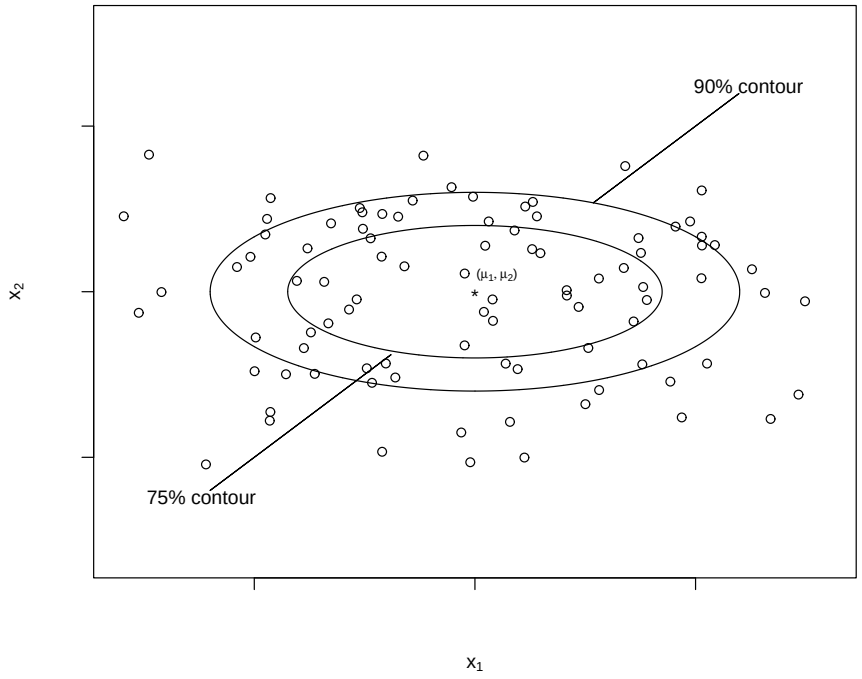


Figure 1.2: Ellipse's delineating regions of equivalent probability.

1.2 Affine Equivariance and Maximum Likelihood Estimation

As a preliminary to a discussion on the affine equivariance of robust, high breakdown estimates for location and scale we introduce the concept of maximum likelihood estimation. In the particular setting where $\mathbf{X}$ has a joint probability density

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

for an observed sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ the maximum likelihood estimate, MLE, for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ is given by those parameters $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ which satisfy

$$L_n(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $L_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the likelihood given for this model by

$$L_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{n} P_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}_i)$$

so that

$$L_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} e^{-1/2 \sum_{i=1}^{n} (\mathbf{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})}$$

On taking logarithms and differentiating $L_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ it can be shown that

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

We say an estimator $T(\mathbf{X})$ of location is affine equivariant if any linear transformation of $\mathbf{X}$ transforms the estimator T likewise,

$$T(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}T(\mathbf{X}) + \mathbf{b}$$

for any $\mathbf{b}$ being a constant vector in $\Re^p$ and $\mathbf{A}$ being any non-singular $p \times p$ constant matrix. For the maximum likelihood estimator $T(\mathbf{X}) = \bar{\mathbf{X}}$ clearly

$$\begin{aligned} T(\mathbf{A}\boldsymbol{X} + \mathbf{b}) &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{A}\mathbf{X}_i + \mathbf{b}) = \mathbf{A}(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i) + \mathbf{b} \\ &= \mathbf{A}T(\mathbf{X}) + \mathbf{b} \end{aligned}$$

Thus the maximum likelihood estimator of location for the normal parametric family is affine equivariant.

Now consider the estimator for scatter. An estimator for scatter $S(\mathbf{X})$ is affine equivariant if and only if

$$S(\mathbf{AX} + \mathbf{b}) = \mathbf{A}S(\mathbf{X})\mathbf{A}^\top$$

for all $\mathbf{A}$ and $\mathbf{b}$ defined as above. Regarding the MLE for scatter,

$$S(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top,$$

we can use the equivariance of $T(\mathbf{X})$ to formulate

$$S(\mathbf{AX} + \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{AX}_i + \mathbf{b} - \mathbf{A}\bar{\mathbf{X}} - \mathbf{b})(\mathbf{AX}_i + \mathbf{b} - \mathbf{A}\bar{\mathbf{X}} - \mathbf{b})^\top$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top\mathbf{A}^\top = \mathbf{A}S(\mathbf{X})\mathbf{A}^\top.$$

Clearly the maximum likelihood estimator of $\hat{\boldsymbol{\Sigma}}$ for the multivariate normal distribution satisfies affine equivariance.

Although equivariant these estimates for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, using our sample data, can be impacted by the very outliers we are trying to detect, diminishing our chances of identifying them. Examining observations for potential outlyingness demands estimates, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, for the centroid and covariance parameters, respectively, necessarily robust to corrupt data.

1.3 M-estimate

Suppose $\boldsymbol{\Sigma}$ was known and we needed to estimate $\boldsymbol{\mu}$. The MLE for $\boldsymbol{\mu}$ can be found by the minimization of

$$\rho_n(\boldsymbol{\mu}) = \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}),$$

ignoring the constant term in the expression for the log-likelihood. This is now equivalent to solving,

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{x}_i - \boldsymbol{\mu}) = 0 \tag{1.2}$$

to obtain an estimate $\hat{\boldsymbol{\mu}}$, where

$$\boldsymbol{\psi}(\mathbf{x}_i - \boldsymbol{\mu}) = \frac{\partial}{\partial \boldsymbol{\mu}} \rho_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$$

Here $\boldsymbol{\psi}$ is a vector function. In discussing generalizations for the choice of $\boldsymbol{\psi}$ we briefly discuss the case of $p = 1$ and $\sigma = 1$ and the theory of M-estimates.

The idea of the M-estimate for location was introduced by Huber (1964) and consisted of the generalization of the maximum likelihood estimator (Hampel et al 1986) which can also be defined as $\hat{T}$ **satisfying**

$$\min_T \left(-\sum_{i=1}^{n} \log f(x_i - T) \right) = -\sum_{i=1}^{n} \log f(x_i - \hat{T}).$$

The generalization is to consider

$$\min_T \sum_{i=1}^{n} \rho(x_i - T) = \sum_{i=1}^{n} \rho(x_i - \hat{T})$$

for a ρ no longer restricted to the negative of the logarithm of the normal density.

The overall performance of an estimator for a parameter is typically measured by its expected loss, such as the Mean Squared Error (MSE). A minimax estimator seeks to *minimize* the *supremum* of this expected loss over a class of ϵ contaminated symmetric distributions (Huber 1964). Huber (1964,1973) proposed using the minimax argumentation for an estimator of symmetric distributions like the normal where

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| < k \\ k|x| - \frac{1}{2}k^2 & \text{for } |x| \geq k \end{cases}$$

which leads to equation (1.2) with

$$\psi(x) = \max(-k, \min(k, x)), \tag{1.3}$$

where k is determined as a function of the sample proportion, ϵ, of *contamination* that yields the minimax solution.

Huber's minimax choice of ψ is then given in Figure 1.3. Note in equation (1.2) with this choice of ψ there is a smooth reigning in of outliers, to give them less weight than those observations composing the bulk of the data.

As an aside, with regard to Huber's minimax, an M-estimate can also be seen as the solution to an M-functional designed to minimize a loss function. If $Y \sim N(\mu, \sigma^2)$ and $Z \sim N(0, s^2)$ then Kozek (2003) shows us that the M-functional coincides with the p-quantile of $V = Y - Z$. It is also noticed in Kozek (2003) that a Huber minimax function, with parameter s, corresponds to a robust estimate of the p-quantile of V, for $p = 0.5$ the median, when Z is uniformly distributed on the interval $[-s, s]$.

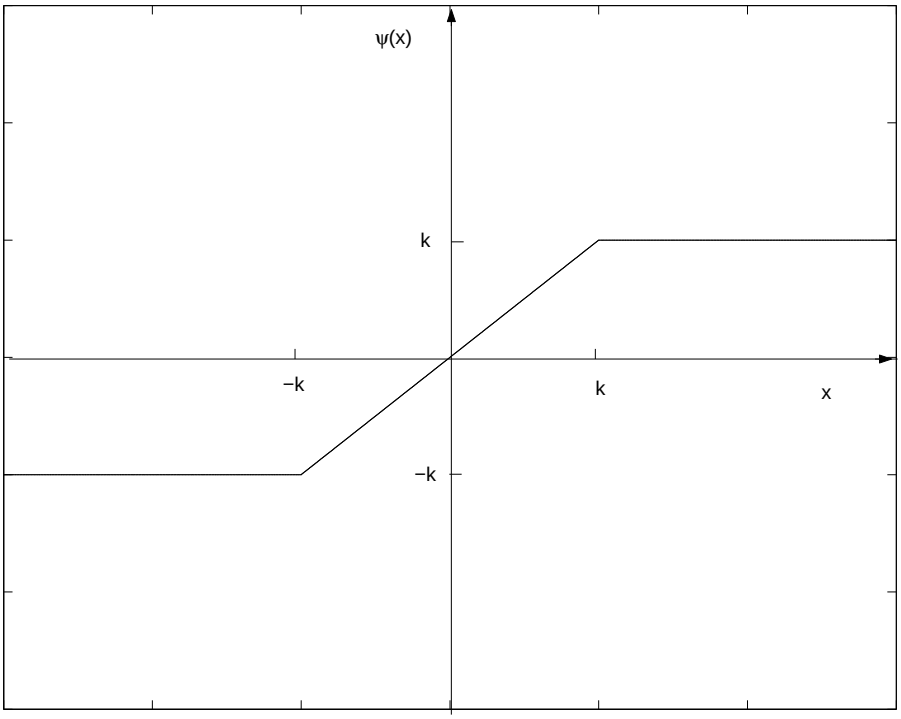


Figure 1.3: Huber Minimax

Huber's minimax ψ inspired Hampel (1968, 1974) to introduce a three-part redescender for ψ which is often quoted as *Hampel's Psi function*, depicted in Figure 1.4. Any M-estimator with a ψ-function which vanishes beyond some central region is termed a *redescending* M-estimator (Huber 1981, Hampel et al 1986) and any observations beyond this region are considered necessarily outlying and disappear. Hampel's Psi function is governed by the

following bounds:

$$\psi(x) = \begin{cases} x & 0 \leq |x| \leq a \\ a \text{ sign}(x) & a \leq |x| \leq b \\ a\frac{c-|x|}{c-b}\text{sign}(x) & b \leq |x| \leq c \\ 0 & c \leq |x| \end{cases} .$$

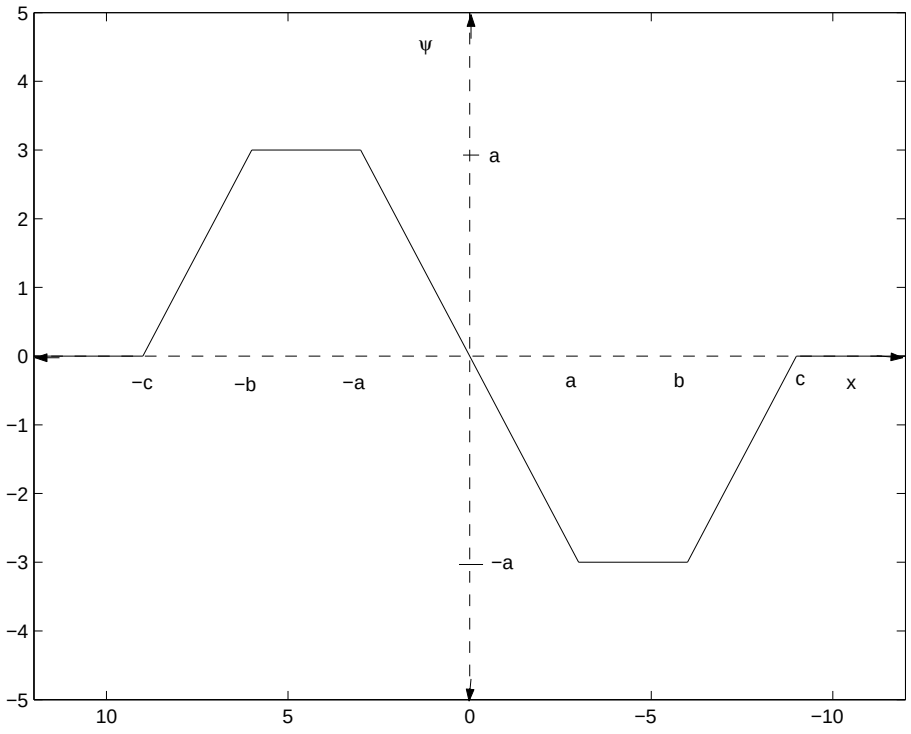


Figure 1.4: Hampel's Psi function

These estimates for location and similar generalizations for scale estimation are not a suitable starting point when we want to identify outliers since the degree of contamination needs to be known precisely beforehand. For example Huber's Proposal 2 (Huber 1981, Clarke and Milne 2004), whereby he finds those estimates for both location and scale simultaneously by maximizing the likelihood, i.e. visualize k as chosen in relation to ϵ, the proportion of contamination, through formula 5.21 in Huber (1981).

Huber (1981) also provides estimates of location and scale for the multivariate case but there does not exist a simple analytic solution for an estimate of location (Hampel et al

1986).

1.4 Robustification of Univariate Regression

The origins of many robustifcation methods for location and scale for multivariate data analysis can be traced back to the robustification of univariate regression analysis. By definition, **the ideal *robust* estimate for any parameter implies any potential *outlying* data has been ignored or weighted accordingly**. In both cases the potential outlier has been detected.

We therefore examine methodology used to robustify univariate regression.

The least squares estimate for the linear model with a single response variable,

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ and $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are p-dimensional vectors of covariates and regression coefficients, is known to be efficient when the fitted data is normally distributed but is not robust to outliers. The least squares estimate possesses the lowest possible breakdown point of $\epsilon^* = 1/n$, where the finite sample breakdown point of an estimator T on the topological space Ω, which for our purposes is the space of samples in $\Re^p$, is defined (Donoho and Huber 1983, Vandev and Neykov 1998)

$$\epsilon_n^*(T) = \tfrac{1}{n}\min\{m : \sup_{\Omega_m} ||T(\Omega_m)|| = +\infty\}.$$

Here Ω_m is any sample obtained from Ω after replacing m points in Ω with arbitrary values. Thus we consider, for various m, the estimate T as being unduly affected by as little as m contaminants and the smallest m/n for which this property holds is the breakdown point.

This susceptibility of the least squares estimate to the impact of even a single outlier motivated the search for more robust methods of regression. To obtain more robust estimates for univariate regression analysis the suggestion was to use a one-step M-estimate (Huber 1973) which is defined

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i/\hat{s}) \tag{1.4}$$

where $r_i = Y_i - \boldsymbol{\beta}^\top \mathbf{x}_i$ and ρ is a symmetric, convex function with a unique minimum at zero. The scale parameter, $\hat{s}$, is introduced to force invariance with respect to a magnification of the error scale.

Putting $\psi = \rho'$ in (1.4) requires a solution to

$$\sum_{i=1}^{n} x_{ij} \psi \left(\frac{Y_i - \boldsymbol{\beta}^\top \mathbf{x}_i}{\hat{s}} \right) = 0, \quad j = 1, \ldots, p,$$

which can be solved for $\boldsymbol{\beta}$ by applying a Newton-Raphson iteration at least once (Huber 1981),

$$T^{(m+1)} = T^{(m)} + \frac{\frac{1}{n} \sum \psi \left(\frac{Y_i - T^{(m)}}{\hat{s}^{(0)}} \right) \hat{s}^{(0)}}{\frac{1}{n} \sum \psi' \left(\frac{Y_i - T^{(m)}}{\hat{s}^{(0)}} \right) \hat{s}^{(0)}}$$

where $T = \boldsymbol{\beta}^\top \mathbf{x}_i$ and preliminary estimates for $\boldsymbol{\beta}$ and the scale parameter $\hat{s}$ have been used. If we can assume the underlying distribution is symmetric and the ψ skew-symmetric, then $T^{(1)}$ is asymptotically $T^{(\infty)}$ so only *one* iteration is required (Huber 1981). The preliminary estimates for $\boldsymbol{\beta}$ and $\hat{s}$ are usually derived from the least squares estimate despite its non-robustness (Huber 1973).

This M-estimate of regression is extremely vulnerable to leverage points, still possessing a breakdown-point of only $\epsilon^* = 1/n$, and so the generalized M-estimators, or GM estimators, were introduced (Mallows 1975) whereby we seek to minimize

$$\textstyle\sum_{i=1}^{n} w(\mathbf{x}_i) \rho(r_i/\hat{s})$$

with a weight function, w, which bounds the influence of any outlying $\mathbf{x}_i$ (Rousseeuw

1984). This procedure was found to have a breakdown-point of at most $1/(p+1)$ (Maronna, Bustos and Yohai 1979) and was therefore superseded by the least median of squares or LMS estimation technique (Rousseeuw 1982,1984, Rousseeuw and Leroy 1987). The LMS is the estimator which minimizes the objective function

$$\text{median}_i \{(r^2)_{i:n}\} \qquad i = 1, 2, \ldots, n$$

where $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \ldots \leq (r^2)_{n:n}$ are the ordered squared residuals and has the maximum possible finite sample breakdown point of $\frac{n-\lfloor \frac{n+p+1}{2} \rfloor}{n}$. Thus with regard to linear regression the LMS is robust to approximately 50% of the data being contaminated, however LMS estimates can be quite unstable (Hettmansperger and Sheather 1992, Shertzer and Prager 2002). Indeed it has been shown (Shertzer and Prager 2002) that although the LMS estimate is robust to outliers, it can be much more sensitive to small data changes than even the classical least squares estimate. Such data changes may occur when adding new observations or shifting an entire data set. The LMS has also been criticized (Hettmansperger and Sheather 1992) for an abnormally slow rate of convergence, $n^{1/3}$, to a non-normal asymptotic distribution and its lack of efficiency when errors are normally distributed (Hampel et al 1986).

1.5 S-estimate

An M-estimate for scale is central to the definition of the S-estimate, which is a variant of the LMS (Hampel et al 1986). It was constructed (Rousseeuw and Yohai 1984) in an attempt to find a regression estimate with high-breakdown and the asymptotic properties of an M-estimate. The formal definition of an S-estimate is to **minimize** $\hat{s}$ **subject to**

$$\frac{1}{n} \sum_{i=1}^{n} \rho(r_i/\hat{s}) = K \tag{1.5}$$

where the constraint K is a tuning constant chosen to reflect the assumed underlying distribution F. An example for K would be to consider $\rho(y) = y^2$ in which case we would be dealing with the usual least squares and $K = 1$ if $\varepsilon_i \sim N(0, 1)$ say, or K=0.5 for maximum breakdown. The S-estimate possesses the high-breakdown and the asymptotic normality we desire of a robust statistic (Rousseeuw and Yohai 1984) and yet it is computation intensive and when a sample is heavily contaminated the S-estimate can yield multiple solutions (Woodruff and Rocke 1994).

Another variant of the LMS estimator (Hampel et al 1986) is the high-breakdown estimator called the least trimmed squares or LTS (Butler 1982, Rousseeuw 1984). This is another type of S-estimator and results in the derivation of a least squares estimate after a pre-specified proportion of the highest squared residuals has been removed from the data, the LTS seeks to minimize,

$$\sum_{i=1}^{h} (r^2)_{i:n} \tag{1.6}$$

where $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \ldots \leq (r^2)_{n:n}$ are the ordered squared residuals and $h = \lfloor \frac{n+p+1}{2} \rfloor$ (Rousseeuw and Leroy 1987). A disadvantage with the LTS estimator is that while it converges in distribution, asymptotically, to a normal distribution, it has only 7.1% asymptotic efficiency at the normal model (Clarke 2000) due to the high proportion of trimming.

1.6 M-estimate for Multivariate Data

The M-estimate was first extended to applications involving multivariate data sets by Maronna (1976). It is expressed more formally (Maronna 1976, Lopuhaa 1989) as solutions to the simultaneous equations

$$\frac{1}{n} \sum_{i=1}^{n} u_1 \left[(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \right]^{1/2} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} u_2\left[(\mathbf{x}_i-\hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}})\right](\mathbf{x}_i-\hat{\boldsymbol{\mu}})(\mathbf{x}_i-\hat{\boldsymbol{\mu}})^\top = 0,$$

thus representing the M-estimate as a weighted mean (Huber 1981) where u_1 and u_2 are real valued functions on $[0,\infty)$, nonincreasing and continuous such as the maximum-likelihood discussed above. The maximum likelihood at the normal distribution is then given by the choice of $u_1(s) = -\frac{1}{s}\frac{d[\log f(s)]}{ds}$ and $u_2(s^2) = u_1(s)$, for $s > 0$.

Although affine equivariant the obstacle with using an M-estimate as the launching platform when trying to identify outliers in multivariate data is its breakdown point (Maronna 1976, Huber 1981, Lopuhaa 1989) of at most $\epsilon^* = \frac{1}{(p+1)}$ for data sets of dimension p. Of course for high-dimension data sets the susceptibility of the M-estimate to only a few outliers corrupting the initial estimate for location becomes unacceptable.

Another pivotal discussion in the search for optimal M-estimates involved using the Influence Function (Hampel 1968, 1974, Hampel et al 1986) where an estimator T for location, assuming the distribution of $\mathbf{X}$ is F, can be evaluated in terms of its influence function. Here the estimator based on a sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ is evaluated in terms of the estimating functional $T[F_n]$ where F_n is the empirical distribution that attributes atomic mass $\frac{1}{n}$ to each point $\mathbf{x}_i$. The influence function is then

$$IF(\mathbf{x},F,T) = \lim_{\epsilon\downarrow 0}\frac{T[(1-\epsilon)F+\epsilon\delta_{\mathbf{x}}]-T[F]}{\epsilon}$$

where $\delta_{\mathbf{x}}$ is the distribution attributing a point mass of one at $\mathbf{x}$.

This function allows us to measure the impact an infinitesimal contaminant at $\mathbf{x}$ would have upon the estimate T. If we replace ϵ with $1/n$ each observation could be considered a suspect outlier and then checked for its impact on our estimate. The disadvantage, for our purposes, is that the Influence Function only tests for *one* outlier (Lopuhaa 1989), this restriction renders the Influence Function impractical since there are potentially $n - \lfloor\frac{n+p+1}{2}\rfloor$ outliers according to the maximum breakdown estimate we are seeking.

1.7 S-estimate for Multivariate data

As referred to earlier, S-estimators were originally devised to robustify regression analysis, as in (1.5). The multivariate equivalent (Lopuhaa 1989) involves minimizing $\mathbf{S}$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho[(\mathbf{x}_i-\hat{\boldsymbol{\mu}})^\top \mathrm{S}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}})]^{1/2}=K. \tag{1.7}$$

When analyzing high-dimension data the classic S-estimator only recognizes, as outliers, points with huge Mahalanobis distances from the sample centroid, distances which occur under the assumption of normality at a rate of much less than 1:1 000 000 000 (Rocke 1996). Changes can be made to the S-estimate to increase *rejection probability* but these are at the cost of efficiency and *gross error sensitivity* (Rocke 1996) where the latter is defined (Huber 1981, Hampel et al 1986) in relation to the influence curve as

$$\gamma^* = \sup_x |IF(x,F,T)|.$$

This is the upper bound of the influence a contaminant can have on an estimate or its maximum bias.

1.8 The MVE and MCD estimates

The S-estimate is a generalization (Woodruff and Rocke 1994, Lopuhaa 1997) of the Minimum Volume Ellipsoid (MVE) (Butler 1982, Rousseeuw 1983, Rousseeuw and Leroy 1987) which is a combinatorial, affine equivariant estimate which is defined as **finding** $\hat{\boldsymbol{\mu}} \in \Re^p$ **and** $\hat{\boldsymbol{\Sigma}} \in PDS(p)$ (**Positive-definite symmetric** $p \times p$ **matrices**) **minimizing the determinant of** $\hat{\boldsymbol{\Sigma}}$ (Lopuhaa and Rousseeuw 1991) **subject to**

$$\#\Big\{i : (\mathbf{x}_i-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}) \le c^2\Big\} \ge \lfloor \tfrac{n+p+1}{2} \rfloor$$

such that from all subsets of size $\lfloor\frac{n+p+1}{2}\rfloor$ from a sample of size n, $\hat{\boldsymbol{\mu}}$ describes the centroid and $\hat{\boldsymbol{\Sigma}}$ the covariance matrix of that subset contained by the smallest ellipsoid. This equates to finding the centroid of the minimal volume ellipsoid covering a subset of at least $h = \lfloor\frac{n+p+1}{2}\rfloor$ points and the most appealing aspect of the MVE is its high breakdown:

$$\epsilon^* = \frac{n - \lfloor\frac{n+p+1}{2}\rfloor}{n}$$

which converges to 1/2 as $n \to \infty$.

If one assumes normality the sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ is governed by an ellipsoidal probability density $\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} f[\{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}^{1/2}]$ and a natural choice for c (Lopuhaa and Rousseeuw 1991) would be the value for which $P_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left\{(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^1(\mathbf{X}-\boldsymbol{\mu}) \leq c^2\right\} = \frac{1}{2}$ such that $c^2 = \chi^2_{0.5,p}$ since $(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_{(p)}$ (Mason and Young 2002).

Rousseeuw (1983), when discussing the asymptotic properties of the MVE, noticed that the distribution of the MVE converges to a normal distribution at an abnormally slow rate, $n^{1/3}$, and so for our purposes cannot be recommended as a suitable starting point for outlier identification techniques (Gather and Becker 1998).

Another approach (Rousseeuw and Leroy 1987) is to apply the Least Trimmed Squares (LTS) estimator, discussed in section 1.5, to multivariate data. This results in the Minimum Covariance Determinant or MCD estimator which has an objective function identical to that of the MVE but converges at a rate of $n^{1/2}$. Instead of finding a minimum ellipsoid the MCD finds the mean of the $\lfloor\frac{n+p+1}{2}\rfloor$ points for which the determinant of the corresponding covariance matrix is minimal, yielding an estimate which has the same breakdown point of $\epsilon^*_n = \frac{\lfloor(n-p+1)/2\rfloor}{n}$ as the MVE.

1.9 Computational Expense

The search for that subset yielding the minimum covariance determinant, or indeed the minimum volume ellipsoid, rapidly approaches the testing of an unacceptably large number of subsets,

$$\frac{n!}{k!(n-k)!},$$

where $k = (\lfloor (n+p+1)/2 \rfloor)$. Some examples of the number of subsets, N, that are of size $\lfloor \frac{n+p+1}{2} \rfloor$, for which determinants need to be checked are shown in Table 1.1 for various sample sizes n of dimension p.

n	p	N
20	2	167960
	7	38760
50	2	1.2155×10^{14}
	10	4.7129×10^{13}
100	2	9.8913×10^{28}
	20	1.3746×10^{28}

Table 1.1: subset count

This poses a computational problem because we need time effective estimation techniques and for data sets even as small as n=100 the amount of time required to check every subset is not plausible. This necessitates the need for algorithms that can provide an *estimate* for the MCD ("the algorithm is the estimator" Woodruff and Rocke 1994).

An algorithm which can, with high probability, yield a close approximation is described below and based on Rousseeuw and Leroy (1987) and Woodruffe and Rocke (1993). This algorithm was used for all of the tests conducted for multivariate analysis in this thesis when using Matlab:

1.10 MCD Algorithm

1. Randomly select $J = p + 1$ points, where p is the dimension of the data, and compute the mean $\hat{\boldsymbol{\mu}}_J$ and the covariance matrix $\hat{\boldsymbol{\Sigma}}_J$ of this subset of J points. If $\hat{\boldsymbol{\Sigma}}_J$ is singular randomly select another subset of J points.

2. Compute the Mahalanobis distances of each n sample points from the centroid of this subset, $\hat{\boldsymbol{\mu}}_{\boldsymbol{J}}$,

$$M_i = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_J)^\top \hat{\boldsymbol{\Sigma}}_J^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_J). \tag{1.8}$$

3. Sort these distances into ascending order and the sample points corresponding to the first $h = \lfloor (n + p + 1)/2 \rfloor$ distances become the new subset.

4. Calculate the Mahalanobis distances of all n sample points from the centroid of this subset then apply step 3.

5. Conduct step 4 two times.

6. Record the mean, covariance matrix and determinant of the final subset obtained.

The above 6 steps are performed k times where

$$k = \frac{\log(0.05)}{\log(1 - \dfrac{\binom{h}{J}}{\binom{n}{J}})}. \tag{1.9}$$

This gives us a 95% chance of selecting J non-contaminated points in the event that $n - h$ points are contaminated which ensures the estimate achieves maximum breakdown.

7. From the resulting k subsets obtained in steps 1 to 6 we select those that correspond

to the 10 smallest determinants.

8. For each of these 10 subsets we apply steps 3 and 4 until convergence.

9. Select the subset possessing the covariance matrix yielding the minimum determinant of these 10 converged to subsets as the chosen MCD estimate of location.

1.11 Outliers

Once we have established a robust estimate for location and scale we are in a position to search for any abnormal deviancy from this centroid estimate with respect to the scatter of the data. The classification of outliers generates four different contaminant types which may require different search methodologies for detection. When confronted with multivariate data there are observations that are relatively easy to identify as outliers such as solitary strays or even a scatter of stray data points, termed linear and radial outliers (Rocke and Woodruff 1993). By simply using a robust estimate of location and scale one can define an *outlier region* using a some pre-specified cut-off value, or *fixed threshold* (Becker and Gather 1999, Rousseeuw and van Zomeren 1990, Rocke and Woodruff 1996, Penny 1994, Hadi 1992 and 1994, Hadi, Jeffrey and Simonoff 1993, Aktinson 1992, Gervini 2003) and simply identify any data points within these extreme regions as outliers.

There are two other types of outlier which can thwart most algorithms successful at detecting the linear and radial contaminants. The first type, *shift outliers*, are those groups of outliers composing clusters which exhibit a similar shape matrix to the majority data but are shifted from the mean of this majority sample population. If the main data is distributed $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a cluster of shifted outliers could be distributed $N(\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{\Sigma})$ for instance, which corrupts the metric based on the Mahalanobis distance (Rocke and Woodruff 1999). The Mahalanobis distance and Hotelling's T^2 are consequently warped and the statistical distances between the outlying and inlying data may disappear. The

magnitude of the displacement, from a population mean, around which any cluster or series of clusters may be centred, often becomes insignificant (Rocke and Woodruff 1999).

The fourth and equally dangerous outlier type is the *point mass* outlier, (Pena and Prieto 2001), which is the result of a high concentration of contaminants within a small region. Such contamination inflates the robust MCD estimate for location and scale into its direction since one is using a metric involving the covariance matrix, thus distorting distances.

The latter two varieties of outlier above can cause serious problems whereby *outlying cases appear inlying* which is understood to be the *masking* effect and *inlying cases appear outlying* which is termed *swamping* (Hawkins 1992).

When dealing with normal data the *outlier* must, by our definition, **be generated according to an alien probability density or distribution $N(\boldsymbol{\mu_c}, \boldsymbol{\Sigma_c})$ with respect to a main population $\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to be identified as a contaminant**. If one is seeking an *outlier region* any single outlier or group thereof must, to be detectable, exhibit a pattern centred about a mean $\boldsymbol{\mu_c}$ displaced from the mean of the majority data $\boldsymbol{\mu}$. Any displaced data may or may not be scattered according to a contaminant variance with respect to that of the main sample population.

Once we have established the robust estimate for location and scale we can go about trying to locate data lying beyond the boundaries of acceptable extremities imposed by these estimates. One naturally assumes, or hopes, that the contaminant distribution is centred about a mean, $\boldsymbol{\mu_c}$, well beyond these boundaries and if this is not the case one may require a non-distance based plan of attack to uncover them. Indeed if a contaminant is *not* outlying, then distance-based outlier detection methodology will necessarily fail to identify them as *outliers* .

1.12 Fixed Threshold Detection Methods

When calculating the Mahalanobis distance

$$M_i = \sqrt{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})}$$

for each point from the estimate of location $\hat{\boldsymbol{\mu}}$, with respect to the estimate of covariance $\hat{\boldsymbol{\Sigma}}$, we can assume the estimates converges to the underlying parameters:

$$M^2 = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_{(p)} \tag{1.10}$$

for p dimensional data (Mason and Young 2002). This being the case we have a fundamental probability measure that can be used to locate extreme observations using distance-based methodology. Seeking a pre-specified level of significance, a fixed threshold defining a cut-off value to an outlier region is the basis of the following three, prevailing, outlier detection algorithms.

1.12.1 Robust fixed threshold

Rousseeuw and van Zomeren (1990) use (1.10) to establish the boundaries for *inlying* data or, equivalently, the *cut-off* values which will denote the beginning of an outlier region. For example in their analysis (Rousseeuw and van Zomeren 1990) of the 3 dimensional Stackloss data (Brownlee 1965) they assert the outlier region as the set of all possible observations which would satisfy

$$M_i > \sqrt{\chi^2_{3,0.975}} = 3.06$$

and identify cases 1,2,3 and 21 as outliers, confirming Rousseeuw and van Zomeren (1990), from the 21 observations composing the Stackloss data set. It is noted that Rousseeuw and van Zomeren (1990) begin with an MVE estimate for location and scale to guard against the impact of up to $\lfloor n - \frac{(n+p+1)}{2} \rfloor$ possible outliers, corresponding with, the highest

possible breakdown. It has also been noted in Atkinson (1982) that these 4 points are not necessarily outlying and in discussions to Atkinson (1982), D.A. Preece, M.A. Aitkin and G.A. Barnard, dismiss this Stackloss data set as being of any use to an analysis because it was not generated by experimental design.

The methodology for outlier detection described in Rousseeuw and van Zomeren (1990) was assessed using a series of Monte Carlo experiments for bivariate data sets $\{\mathbf{X}_1, \mathbf{X}_2\}$, each variable distributed $N(0,1)$. The results are displayed in Table 1.2 for sample sizes $n = 20, 50, 100$. Each sample of size n had a proportion, ϵ, of the variable $\mathbf{X}_2$ generated $N(d,1)$, thus shifted from the majority sample mean of zero. The Monte Carlo samples were separated into 5 types depending on the proportion of the $\mathbf{X}_2$ contamination and its severity:

• Proportion of contamination $\epsilon = 0$ which corresponds to the assessment of *clean* data sets.

• $\epsilon = 1/n$ of $\mathbf{X}_2 \sim N(5.4324, 1)$.

• $\epsilon = 1/n$ of $\mathbf{X}_2 \sim N(10.8648, 1)$.

• $\epsilon = 0.3$ of $\mathbf{X}_2 \sim N(5.4324, 1)$.

• $\epsilon = 0.3$ of $\mathbf{X}_2 \sim N(10.8648, 1)$.

Noting that 5.4324 and 10.9648 correspond to $d = 2\sqrt{\chi^2_{0.975,2}}$ and $4\sqrt{\chi^2_{0.975,2}}$ respectively (Juan and Prieto 2001).

Figures 1.5-1.8 show examples of such contamination for samples of size $n = 100$. For the smaller size displacement of outlying data, $d = 2\sqrt{\chi^2_{0.975,2}}$, there is evidence of a likely uncertainty creeping in as to whether the contaminants are *outlying enough* to warrant *outlier status*. For the larger, more definitive level of displaced outlier mean, we can see there should be no uncertainty although *good* data can sometimes confuse the issue because a *less than significant* percentage of good data may be expected to be *outlying*.

For this method, Rousseeuw and van Zomeren (1990), the MVE estimate for location and scale is found and then any observation possessing a Mahalanobis distance

$$M_i > \sqrt{\chi^2_{0.975,p}}$$

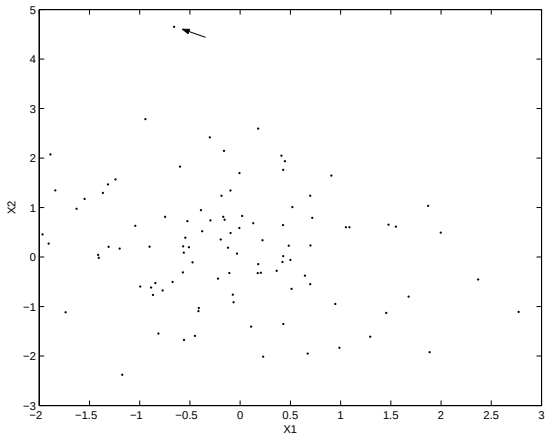


Figure 1.5: Single outlier displaced $d = 2\sqrt{\chi^2_{0.975,2}}$.

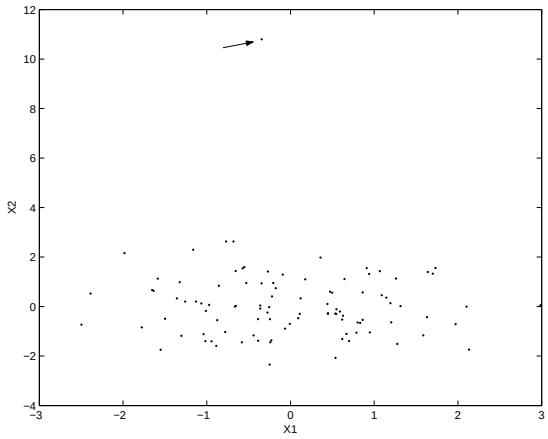


Figure 1.6: Single outlier displaced $d = 4\sqrt{\chi^2_{0.975,2}}$.

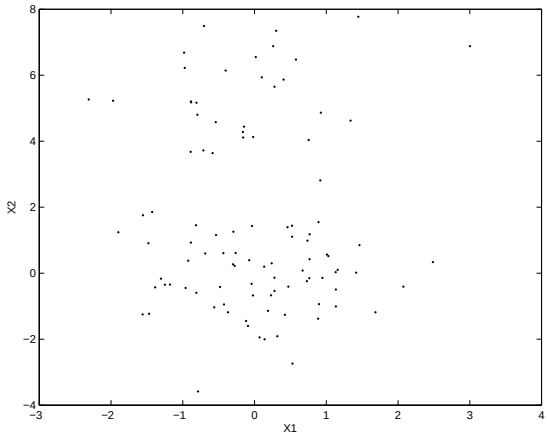


Figure 1.7: Thirty outliers displaced about a mean $d = 2\sqrt{\chi^2_{0.975,2}}$ from underlying centroid.

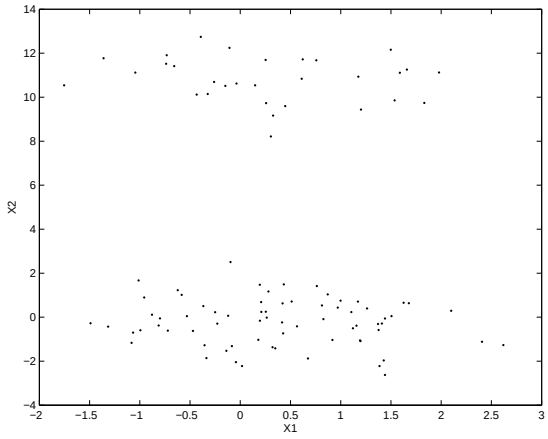


Figure 1.8: Thirty outliers displaced about a mean $d = 4\sqrt{\chi^2_{0.975,2}}$ from underlying centroid.

is identified as an outlier. The power p_t of this algorithm, defined as the proportion of each of the 1000 generated samples in which outliers were identified, was calculated and tabulated in Table 1.2 The average proportion of outliers identified over all samples, equivalently the average amount of trimming, $\bar{\alpha}$, advised by their algorithm is also tabulated in Table 1.2.

It is evident from Table 1.2 that this methodology is too sensitive with a high proportion of clean data sets containing observations identified as outliers. When encountering data sets possessing solitary strays, this algorithm routinely over trims the data set which yields unnecessary loss of information.

1.12.2 Forward Search

Hadi (1992, 1994) reminds us that an observation with a large Mahalanobis distance from a centroid may not necessarily be outlying. Due to the impact of *swamping* a small cluster of outliers may inflate the estimate for the Covariance Matrix, $\hat{\boldsymbol{\Sigma}}$ and attract the location estimate away from the centroid of the majority data, $\hat{\boldsymbol{\mu}}$. This may result in clean data exhibiting outlyingness since its Mahalanobis distance may be large with respect to these *corrupt* estimates. Hadi (1992, 1994) warns of *masking* also, whereby those observations comprising this outlying cluster possess small Mahalanobis distances.

To counter this Hadi (1992, 1994) devised a Forward Search algorithm to detect outliers in multivariate data sets of dimension p. Initially the algorithm orders the sample in ascending order according to corresponding Mahalanobis distance from a robust estimate for location with respect to a robust estimate for scale. Hadi then divides this ordered data set into two subsets, the first, called the basic subset, containing the closest $p+1$ observations to this estimate for $\hat{\boldsymbol{\mu}}$ and the other remaining $n-p-1$ observations. With respect to the estimate for centroid and covariance matrix of this basic subset the whole data set is ordered again, in ascending order of corresponding Mahalanobis distance. At this stage, and each subsequent repetition thereof, it is imperative to note that some members of this basic subset can interchange with complement members due to this re-ordering, therefore some original $p+1$ observations may no longer be a member of this subset, (Atkinson, Riani and Cerioli 2004). After this re-ordering, the $p+2$ observations with the smallest Mahalanobis distance are selected to form a newly *inflated* basic subset. The centroid and covariance matrix of this new basic subset of size $n = p+2$ is used to

re-order the whole data set anew. Observations are continually added to the basic subset, in this way, until a certain stopping criterion is met (Hadi 1992) for a basic subset size of

$$h \geq \lfloor \frac{n+p+1}{2} \rfloor \tag{1.11}$$

observations (Hadi 1994). Equality in (1.11) would correspond to the highest possible breakdown if this subset was used for parameter estimation. Hadi's algorithm continues to inflate the basic subset to contain the entire data set if the stopping criteria is not met. This stopping criteria is satisfied when the Mahalanobis distance, M_i, of the closest observation to any basic subset of size h satisfies

$$M_i > \sqrt{\chi^2_{1-\alpha/n,p}}. \tag{1.12}$$

If (1.12) is satisfied we identify all observations not a member of this basic subset as outliers. If (1.12) is not satisfied, the basic subset is incremented with this closest observation.

A correction factor, based on a large simulation study (Hadi 1994),

$$c_{np} = (1 + \frac{2}{n-1-3p} + \frac{p+1}{n-p})$$

is applied to the computation of the Mahalanobis distances calculated when locating a possible outlier region.

The initial robust estimate for location and scale was computed by Hadi (1992) using the co-ordinate medians as a preliminary estimate for location from which an estimate for scale was derived. This estimate was revised by calculating the centroid and corresponding scale for those $h = \lfloor \frac{n+p+1}{2} \rfloor$ observations closest to this preliminary estimate.

Table 1.3 contains the results using the outlier detection strategy described by Hadi (1992, 1994) and it can be seen that it is a great improvement on the method described in Rousseeuw and van Zomeren (1990). Table 1.3 contains the simulation results of Hadi's algorithm applied to data set scenarios akin to those assessed when using the algorithm devised for the Rousseeuw and van Zomeren (1990) paper. For the larger *outlier mean*, data distributed $N(4\sqrt{\chi^2_{0.975,2}}, \mathbf{I_2})$, the Hadi Algorithm identified close to the correct outlier proportion in more than 99.9% of Monte Carlo samples. Hadi's method, however,

becomes over sensitive as the sample size increases, for example Hadi's algorithm identified outliers in nearly 20% of the clean data sets generated of size $n = 100$. The average amount of trimming is an important statistic to consider and Hadi's algorithm appears to be very strong from this perspective with the figures showing that the trimming amounts are, if not exact, in the vicinity of the exact contamination levels, $\epsilon = 1/n, 0.3$ respectively.

1.12.3 Standardized distances and simulations

An algorithm devised for the estimation of an outlier region was also taken up by Rocke and Woodruffe (1996) starting with a robust estimate for location, $\hat{\boldsymbol{\mu}}$, and scale, $\hat{\boldsymbol{\Sigma}}$, using the Minimum Covariance Determinant (MCD). They standardize the MCD estimate for the scale matrix such that the hth ordered Mahalanobis distance is equal to $\chi^2_{h/n,p}$, where p signifies the dimension of the data set and $h = \lfloor \frac{n+p+1}{2} \rfloor$. Next, using simulations, they establish a cutoff value whereby a pre-specified fraction, α_1, of points on average lie beyond a particular value. Using this cutoff point to establish an inlying region for all data sets of corresponding size and dimension a new covariance matrix is derived using only those observations of the data set which lie within this region. The new location estimate is now the mean of these observations considered inlying. Finally they identify as an outlier any observation whose location with respect to this revised estimate for location and scale is larger than $\chi^2_{1-\alpha_2,p}$ where α_2 is arbitrary and so for simplicity we put $\alpha_1 = \alpha_2$.

The results tabulated in Table 1.4 confirm this algorithm's extremely high sensitivity given the cutoff points established via simulation for these types of data sets. For samples of size $n = 100$ this algorithm was identifying outliers in over 90% of clean data sets. Another drawback with this algorithm is the need to derive cut-off values using simulations, we would like an algorithm that can determine its own cut-off values, an *adaptive* algorithm.

n	ϵ	d	p_t	$\bar{\alpha}$
20	0		0.8888	0.1675
	0.05	5.4324	0.997	0.1674
		10.8648	> 0.999	0.1680
	0.3	5.4324	0.991	0.2879
		10.8648	> 0.999	0.3141
50	0		0.948	0.0898
	0.02	5.4324	0.999	0.0966
		10.8648	> 0.999	0.0945
	0.3	5.4324	> 0.999	0.3028
		10.8648	> 0.999	0.3162
100	0		0.984	0.0598
	0.01	5.4324	> 0.999	0.0673
		10.8648	> 0.999	0.0668
	0.3	5.4324	> 0.999	0.3090
		10.8648	> 0.999	0.3180

Table 1.2: Results of simulations using Rousseeuw and van Zomeren (1990) algorithm.

n	ϵ	d	p_t	$\bar{\alpha}$
20	0		0.051	0.1933
	0.05	5.4324	0.827	0.0629
		10.8648	> 0.999	0.0604
	0.3	5.4324	0.574	0.2958
		10.8648	> 0.999	0.3074
50	0		0.105	0.0244
	0.02	5.4324	0.959	0.0221
		10.8648	> 0.999	0.0222
	0.3	5.4324	0.748	0.2702
		10.8648	> 0.999	0.3033
100	0		0.199	0.0115
	0.01	5.4324	0.972	0.0121
		10.8648	> 0.999	0.0122
	0.3	5.4324	0.869	0.2420
		10.8648	> 0.999	0.3022

Table 1.3: Results of simulations using Hadi (1992,1994) algorithm.

1.13 Cluster Techniques

As mentioned above, a group or cluster of shift observations comprising the same shape as the majority population can be difficult to detect as outlying. The cluster of outliers is even more of a problem when one considers the original robust MCD estimate, itself heuristically calculated since complete enumeration involves impossible computational time scales. The larger the contaminant cluster, in proportion to the majority population, the more likely contaminant data will be included in the final subset arrived at by the MCD algorithm. With these issues in mind we begin the investigation into algorithms designed for cluster detection.

The most basic cluster detection algorithms K-means (Steinhaus 1956-57, MacQueen 1967) and agglomerative hierarchical (Mardia, Kent and Bibby 1979) were examined and then we assessed those procedures outlined in Coleman and Woodruff (2000) when used in conjunction with those of Rocke and Woodruff (1999).

1.13.1 K-means

The K-means clustering algorithm (MacQueen 1967) is an iterative algorithm that minimizes the sum of the distances squared from each observation to its cluster centroid, partitioning n points into k disjoint subsets S_i so as to minimize the sum of squares criterion

$$\omega^2(S) = \sum_{i=1}^{k} \int_{S_i} |\mathbf{x} - \hat{\boldsymbol{\mu}}_i|^2 dp(\mathbf{x}).$$

Here $p(\mathbf{x})$ is the probability mass function of the population, $\hat{\boldsymbol{\mu}}_i$ is the centroid of subset S_i and ensures that for sample data we seek to minimize

$$\omega^2(S) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} |\mathbf{x} - \hat{\boldsymbol{\mu}}_i|^2$$

which is essentially an optimizing partition technique. A predetermined number of clusters,

say K, is assumed to represent the data set and this number is used to specify the number of *seed points* inserted into the domain. A seed point acts as the centroid of a cluster which, at this stage, contains no observations as members. The next step is to cycle through each observation, placing an observation in a cluster corresponding to the seed point that it is closest to using Euclidean distance as the measure of proximity. The locations of the cluster centroids are now recomputed using the points that have merged with each cluster. The next step is to cycle through all of the observations again and because the centroids will alter position, it may be possible for certain observations to be closer to a cluster centroid in a cluster that it is not a member of. This second sweep through the data acts to refine the first sweep by placing observations in clusters centred about that centroid it is closest to. The centroids are updated as new members leave or join the cluster. This refinement is done until no single observation changes clusters in one sweep (Arnott and Evans 2003). Caution is needed as the K-means procedure just outlined will enforce a level of clustering, K, upon the data even if the data is classically uni-modal. It is necessary therefore to determine whether or not these, now exclusive, subsets of the data really do represent observations clustered about shifted means. A preliminary examination of *silhouettes* (Rousseeuw 1987) is used here for the verification of the existence of clusters (Struyf, Hubert and Rousseeuw 1997). Silhouettes is a measure of the partitioning imposed by any configuration of clusters and is formulated by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance from the i-th point to all the other points in its cluster and $b(i)$ is the average distance from the i-th point to all the points in the nearest neighbouring cluster (see Rousseeuw 1987). The number of possible clusters is established by that number yielding the maximum mean silhouette value,

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(i).$$

The maximum $\bar{s}$ is crucial for this analysis as it can indicate the optimal number of clusters possibly present in the data and can be used to confirm the possibility that the data set is best represented without being grouped into clusters. A trial of 10 000 normally

distributed bivariate data sets without contaminants was used to designate a silhouette cut-off value whereby if the maximum mean silhouette value exceeds this value then the corresponding amount of clustering is a valid representation of the sample data. Table 1.5 contains these cut-off silhouette values, c, for sample sizes $n = 20, 50, 100$ of dimension $p = 2$. It shows, for example, that if a silhouette value $c = 0.725$ was used for bivariate samples, of size $n = 20$, the proportion of uni-modal samples clustered in error was less than 5%.

Table 1.6 contains the results using K-means analysis for the type of data sets examined previously using the distance based detection methods already discussed. The figures tabulated refer to the power of K-means to identify any proportion of the planted outliers given cluster analysis was deemed necessary by the associated silhouette value $s > c$. An example for this is using the first result in Table 1.6 where with only a probability of 0.4970 will the silhouette value exceed the relevant cut-off, the average amount of trimming enforced, when the silhouette cut-off value was exceeded, was 0.0698. Note $\bar{\alpha}$ was calculated only for those instances when the silhouette value was exceeded. Also it must be noted that the poor performance of K-means, when used to identify *radial* outliers, is expected since it is designed to locate clusters. The success at locating the stray outlier was only examined to confirm that cluster analysis can never be an all round tool for outlier identification. The K-means is shown to be excellent at identifying outlying clusters for multivariate data when the average outlier displacement was large, say $d = 4\sqrt{\chi^2_{0.975,2}} = 10.8648$, and even for small sample sizes, $n = 20$, was proficient at identifying clustered outliers. For solitary outliers the figures are weak to very poor and with increasing sample size the silhouette value was exceeded less and less.

1.13.2 Agglomerative Hierarchical

The agglomerative hierarchical cluster method was examined here for bivariate data sets whence the interpoint Mahalanobis distances are used to measure the closeness between

n	ϵ	d	p_t	$\bar{\alpha}$
20	0		0.152	0.0213
	0.05	5.4324	0.817	0.0678
		10.8648	> 0.999	0.0651
	0.3	5.4324	0.711	0.2650
		10.8648	0.998	0.3032
50	0		0.752	0.0450
	0.02	5.4324	0.992	0.0664
		10.8648	> 0.999	0.0669
	0.3	5.4324	0.983	0.2970
		10.8648	> 0.999	0.3287
100	0		0.963	0.0468
	0.01	5.4324	> 0.999	0.0551
		10.8648	> 0.999	0.0571
	0.3	5.4324	> 0.999	0.3270
		10.8648	> 0.999	0.3422

Table 1.4: Results of simulations using Rocke and Woodruff (1996) algorithm

n	c
20	0.725
50	0.625
100	0.580

Table 1.5: Silhouette cutoffs for K-means.

n	ϵ	d	$s > c$	$\bar{\alpha}$
20	0.05	5.4324	0.497	0.0698
		10.8648	0.9719	0.0531
	0.3	5.4324	0.9168	0.3038
		10.8648	> 0.9999	0.3000
50	0.02	5.4324	0.0377	0.0281
		10.8648	0.5823	0.0751
	0.3	5.4324	0.8402	0.3020
		10.8648	> 0.9999	0.3000
100	0.01	5.4324	0.0385	0.01654
		10.8648	0.0396	0.03971
	0.3	5.4324	0.7191	0.30168
		10.8648	> 0.9999	0.3000

Table 1.6: Simulation results using K-means.

neighbouring data points. The single linkage method measures the distance between clusters by the distance between the two closest points within the clusters. This agglomerative technique begins with each of the n points belonging to its own cluster, $\mathrm{C}_1, \ldots, \mathrm{C}_n$, together resulting in $\frac{1}{2}n(n-1)$ interpoint standardized Euclidean distances $D_{ij}(i \neq j)$,

$$D_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j),$$

where $\mathbf{S}$ is the diagonal matrix extracted from the covariance matrix of the sample data.

These distances are arranged in ascending order and those points, r and s say, satisfying $D_{rs} = \min(D_{ij})$ are merged to form a new cluster $\mathrm{C}_r + \mathrm{C}_s$, so we now have $n-1$ clusters. This process is repeated for all $\frac{1}{2}n(n-1)$ interpoint distances, clusters being formed, or increasing in size, due to the *agglomeration* of points satisfying

$$\min(D_{ij}). \tag{1.13}$$

If any two points satisfying (1.13) belong to different clusters, then the two clusters the points belong to are merged to form one cluster. Since this procedure continues until there is only one cluster containing all n points it is necessary to establish beforehand if cluster analysis is a feasible representation of the data set. Hierarchical clustering algorithms impose a hierarchical structure upon the data which can be displayed in a dendrogram (Mardia, Kent and Bibby 1979). The correlation between the dendrogram and the $\frac{1}{2}n(n-1)$ interpoint distances can be calculated and is expressed as the Cophenetic Correlation Coefficient,

$$\rho_{\text{cophenet}} = \frac{\sum_{i<j}(Y_{ij} - \bar{Y})(Z_{ij} - \bar{Z})}{\sqrt{\sum_{i<j}(Y_{ij} - \bar{Y})^2 \sum_{i<j}(Z_{ij} - \bar{Z})^2}}.$$

Here Y_{ij} corresponds with each $D_{ij}(i \neq j)$ and Z_{ij} corresponds with the linkage distances between paired objects of neighbouring clusters. This coefficient ρ_{cophenet} can be used to determine if cluster analysis is warranted.

The single linkage technique is problematic according to Wilks (1995) who describes how this clustering method is prone to *chaining*, whereby very large, unrepresentative clusters

are created because of the nearness of points to one side of a cluster. Even if the majority of points are a long distance away from each other, only one, close together pair of points is necessary to cause the two clusters to merge, this method is therefore not a very popular one for use with multivariate data sets.

The single linkage, or nearest neighbour model, was compared with the *complete* linkage method for these tests (Mardia, Kent and Bibby 1979). Since the single linkage model is susceptible to chaining, which can inevitably merge clusters that should remain distinct, the complete linkage algorithm was also assessed for comparison. The complete linkage algorithm only joins an observation to an already existing cluster when it is relatively close to *all* the points in the cluster. This ensures a more definitive clustering and greater sensitivity to data abnormalities. The associated cophnetic matrix (Mardia, Kent and Bibby 1979) and silhouette values (Rousseeuw 1987), were also compared for levels of efficiency.

The results were indeed surprising for the hierarchical methods discussed. It was discovered that the single linkage model performed much better than complete linkage as the latter nearly always grouped the data set into too many clusters. Table 1.7 summarises these findings whereby each models power at correctly identifying the clusters, P_c, is contained. For sample sizes of $n = 20$ for instance, designed with 14 observations distributed $N(0, 1)$ and 6 observations distributed $N(10.8648, 1)$, the single linkage model picked out the two clusters in more than 80% of samples, $P_c = 0.8182$, in comparison with complete linkage, $P_c = 0.0985$, a success rate of less than 10%. A statistic of much greater importance is all those obtained for data composed of 3 clusters which suggests that for data sets of dimension $p > 2$ hierarchical methods are not recommended.

An inspection of Silhouettes in comparison to Cophenetic Correlation Coefficient was conducted when outliers were displaced about a shifted mean of $d = 2\sqrt{\chi^2_{0.975}} = 5.4324$. This resulted in the Silhouette value exceeding its corresponding cut-off on more than double the occasions the Cophenet Correlation Coefficient did.

Table 1.8 gives the results of applying Agglomerative Hierarchical methods using the

single linkage model to bivariate data sets. Any arbitrary set only deemed to be clustered when its consequent Silhouette value had exceeded a cutoff derived using simulations. For example, observe the figures for sample sizes of $n = 20$ where clusters of contaminants, of sample proportion $\epsilon = 0.3$, are centred about a displaced mean $d = 2\sqrt{\chi^2_{0.975}} = 5.4324$. The Silhouette value was found to have an estimated chance of exceeding cut-off of only 0.5929. Of even greater significance is the fact that this hierarchical strategy identifies the radial outlier with a much higher frequency that it does the clustered outliers.

1.13.3 EM-Algorithm

The final cluster methodology to be inspected for bivariate samples involves the Expectation-Maximization Algorithm (Dempster, Laird and Rubin 1977). This algorithm was originally devised to find the maximum-likelihood estimates for the parameters of an underlying distribution for incomplete data-sets. Here we begin with the assumption that some observed data $\mathbf{X}$ is generated by a particular distribution F governed by parameters $\mathbf{\Theta}$. We next assume that a complete data set exists, $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{Y}$ represents the vector of unknown values. In assuming there is a relationship between the observed and unobserved values we also assume the joint probability density,

$$p(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta}) = p(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta})p(\mathbf{X}|\mathbf{\Theta}). \tag{1.14}$$

From (1.14) a new likelihood function can be defined,

$$L(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y}) = p(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta})$$

and since $\mathbf{Y}$ is unknown we can think of $L(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})$ as some function with constants $\mathbf{X}$ and $\mathbf{\Theta}$ whilst $\mathbf{Y}$ is a random variable (Bilmes 1998). The EM algorithm begins with an initial step of estimating the values for $\mathbf{\Theta}$ obtained from the observed data $\mathbf{X}$ and uses these for the expectation step $\mathbf{E}$:

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{(t-1)}) = E[\log p(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta})|\mathbf{X}, \mathbf{\Theta}^{(t-1)}]$$

model	n	number of clusters	P_c
single linkage	20	2	0.8182
		3	0.2452
	50	2	0.8747
		3	0.3892
	100	2	0.8933
		3	0.5631
complete linkage	20	2	0.0985
		3	0.0574
	50	2	0.0367
		3	0.1644
	100	2	0.0284
		3	0.2878

Table 1.7: Single linkage vs complete linkage cluster identification.

n	ϵ	d	$s > c$	$\bar{\alpha}$
20	0.05	5.4324	0.8921	0.0454
		10.8648	> 0.9999	0.0500
	0.3	5.4324	0.5929	0.1786
		10.8648	0.9734	0.2935
50	0.02	5.4324	0.9173	0.00192
		10.8648	> 0.9999	0.0200
	0.3	5.4324	0.389	0.0508
		10.8648	0.9833	0.2939
100	0.01	5.4324	0.9187	0.0094
		10.8648	> 0.9999	0.0100
	0.3	5.4324	0.2066	0.0621
		10.8648	0.9855	0.2959

Table 1.8: Simulation results using the Agglomerative Hierarchical single linkage algorithm.

where $\boldsymbol{\Theta}^{(t-1)}$ initially represents the initial parameter estimates and the *previous* estimate in subsequent iterations. The maximization step **M** is the maximizing of the expectation computed in the previous step (Bilmes 1998),

$$\boldsymbol{\Theta}^t = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t-1)}).$$

The two steps, **E** and **M**, when iterated increase the log-likelihood of the expectation which in turn also increases $\log p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta})$ (Bilmes 1998). For example, to summarize the EM algorithm, one uses estimates for the mean, $\hat{\boldsymbol{\mu}}$, and variance, $\hat{\boldsymbol{\Sigma}}$, of a data set obtained from the observed values and uses these to estimate the expected value of the *missing* values, this is the **E**-step. This estimate for the *missing* values is then incorporated into a transitional *complete* data set to find the maximum likelihood estimates for a new $\boldsymbol{\Theta}$, this is the **M**-step. These updated estimates for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, are then used as the revised input estimates for $\boldsymbol{\Theta}$ in the **E**-step and so on.

The EM algorithm is widely used for deciphering mixture models where the unknown variable or *missing data* is construed as the probability any single observation was generated by a distribution $F(\mathbf{X}|\boldsymbol{\Theta})$ which, once determined, can identify the number of densities responsible for the data. Given a p-dimensional data set of n observations, $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, we can assume they were generated by G number of different Gaussians, for our purposes and without any loss of generality, governed by parameters

$$\boldsymbol{\Theta} = (\mathrm{a}_1, \ldots, \mathrm{a}_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G),$$

where a_i are the mixing coefficients such that

$$\sum_{i=1}^{G} \mathrm{a}_i = 1, \quad \mathrm{a}_i \geq 0.$$

This constraint is imposed to ensure that the sum of the proportions of data, $\mathrm{a}_1, \ldots, \mathrm{a}_G$, generated by each of the G mixture densities is one (Bilmes 1998).

If we assume G is known, the log-likelihood for this density is therefore

$$\log(L(\boldsymbol{\Theta}|\mathbf{X})) = \log \prod_{i=1}^{n} \prod_{j=1}^{G} p(\mathbf{x}_i|\boldsymbol{\theta}_j)^{\ell_{ij}} = \sum_{i=1}^{n} \log \left(\sum_{j \in J_1,\ldots,J_G} \mathrm{a}_j p_j(\mathbf{x}_i|\boldsymbol{\theta}_j) \right) \tag{1.15}$$

where

$$\ell_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to } j\text{th mixture density} \\ 0 & \text{otherwise} \end{cases}$$

and $\boldsymbol{\theta}_j$ represents the parameter values $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G)$ and $J_1, \ldots, J_G$ is a partition of the n observations constituting the data set (Rocke and Woodruff 1999), each partitioned set generated by an independent density. If we introduce an unobserved value $\mathbf{Y} = \{y_{1j}, \ldots, y_{nj}\}$, such that $y_{ij} = j$ if and only if $\mathbf{x}_i$ belongs to the jth mixture density, we arrive at a log-likelihood (Bilmes 1998)

$$\log(p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta})) = \log \prod_{i=1}^{n} (\mathrm{a}_{y_{ij}} p_{y_{ij}}(\mathbf{x}_i|\boldsymbol{\theta}_{y_{ij}})) = \log \sum_{i=1}^{n} (\mathrm{a}_{y_{ij}} p_{y_{ij}}(x_i|\boldsymbol{\theta}_{y_{ij}})) \tag{1.16}$$

where $\boldsymbol{\theta}_{y_{ij}}$ is that particular $\boldsymbol{\theta}_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ which generated the $\mathbf{x}_i$.

Since $\mathbf{Y}$ is a random vector we can use Bayes's rule with the insertion of some preliminary estimates for $\boldsymbol{\Theta} = (\hat{\mathrm{a}}_1, \ldots, \hat{\mathrm{a}}_G, \hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_G, \hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_G)$, say $\boldsymbol{\Theta}^{(i-1)} = \boldsymbol{\Theta}^{(0)}$, since $i = 1$ for the first **E**-step to yield an expected value for this l_{ij},

$$p(y_{ij}|\mathbf{x}_i, \boldsymbol{\Theta}^{(0)}) = \frac{\hat{\mathrm{a}}^{(0)}_{y_{ij}} p_{y_{ij}}(\mathbf{x}_i|\boldsymbol{\theta}^{(0)}_{y_{ij}})}{p(\mathbf{x}_i|\boldsymbol{\Theta}^{(0)})} = \frac{\hat{\mathrm{a}}^{(0)}_{y_{ij}} p_{y_{ij}}(\mathbf{x}_i|\boldsymbol{\theta}^{(0)}_{y_{ij}})}{\sum_{j=1}^{G} \hat{\mathrm{a}}^{(0)}_j p_j(\mathbf{x}_i|\boldsymbol{\theta}^{(0)}_j)} = E(\ell_{ij}) \tag{1.17}$$

which again is the probability that observation $\mathbf{x}_i$ was generated by the jth density.

Once we have established the expected value of every possible l_{ij}, given some $\boldsymbol{\Theta}$, we can use this value in the **M**-step to update the preliminary $\boldsymbol{\Theta}$. This consists of maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$ subject to the constraint $\sum_{j=1}^{G} \hat{\mathrm{a}}_j = 1$ which entails the **M**-step comprise of 3-substeps (Mitchell 1997):

$$\hat{\mathrm{a}}_j^{t+1} = \frac{1}{n} \sum_{i=1}^{n} E(\ell_{ij})$$

$$\hat{\mu}_j^{t+1} = \frac{1}{n\hat{\mathrm{a}}_j^{t+1}} \sum_{i=1}^{n} E(\ell_{ij}) \mathbf{x}_i \tag{1.18}$$

$$\hat{\sum}_j^{t+1} = \frac{1}{n\hat{a}_j^{t+1}} \sum_{i=1}^{n} E(\ell_{ij})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{t+1})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{t+1})^\top$$

The above **E**-step and **M**-steps are iterated until $\boldsymbol{\Theta}^{t+1}$ converges to $\boldsymbol{\Theta}^t$.

The above algorithm is used to establish the different densities responsible for mixture models which can be seen as the parameters for *clustered* data. This algorithm can be used to assign each observation to its respective cluster *if* the data is clustered. If the data is not clustered this algorithm can assign all the observations to the one cluster. A possible drawback is the arbitrary starting points one needs for the preliminary *guessed* values of the parameters $\boldsymbol{\Theta}$ in the first iteration. We can optimize this by using the K-means algorithm, see the EMCD algorithm in Coleman and Woodruff (2000), and it is interesting to notice that equations (1.17) and (1.18) actually compose the K-means algorithm (Mitchell 1997). The next problem area could be the susceptibility of small clusters to yielding singularities which can be taken care of by simply eliminating any such clusters arrived at, at any stage of the algorithm, see the MINO algorithm in Rocke and Woodruff (1999). So to summarize the simulation procedure that follows we use the K-means estimate of *five* clusters for sample sizes of say $n = 100$ which will enforce a clustering of the sample into 5 clusters. After eliminating any clusters of size $< (p + 1)$, where p is the dimension of the data, we proceed with the EM algorithm as described above. It is very often the case that when the iterations have converged the sample will be found assigned to only 1 or 2 clusters *unless* the sample is indeed designed to be composed of more than 2 clusters. Outlier status will be attributed to any observation not belonging to the largest cluster. When applying this algorithm of samples of size $n = 20$ and $n = 50$ the initial K-means step was used to sort the samples into 3 and 4 clusters respectively. The K-means was limited to finding this number of clusters to prevent the frequency of singular covariance matrices occurring.

Table 1.9 contains the results for the application of this algorithm to data sets generated identically to the scenarios examined above. The advantage of using an algorithm for cluster analysis that is more complex than, say the K-means or Agglomerative Hierarchical

methodologies assessed above, has not really resulted in a better outcome. It is evident from the figures in Table 1.9 that there was a low rate, p_t, of abnormality detection and when clusters were detected the proportion, $\bar{\alpha}$, of *outliers* composing the necessarily smaller clusters was too high. For example, when samples consisted of one outlier, centred about a displaced mean $d = 4\sqrt{\chi^2_{0.975,2}} = 10.8648$, the average trimming advised was more than 30% of the data. This is clearly unacceptable.

n	ϵ	d	p_t	$\bar{\alpha}$
20	0		0.3461	0.25745
	0.05	5.4324	0.2388	0.29365
		10.8648	0.3234	0.30405
	0.3	5.4324	0.4068	0.43145
		10.8648	0.4086	0.47335
50	0		0.6455	0.09046
	0.02	5.4324	0.3756	0.08938
		10.8648	0.7044	0.08216
	0.3	5.4324	0.3699	0.1012
		10.8648	0.5539	0.45852
100	0		0.8153	0.01866
	0.01	5.4324	0.3331	0.03677
		10.8648	0.7862	0.02817
	0.3	5.4324	0.5578	0.33659
		10.8648	0.5269	0.47226

Table 1.9: K-means + MINO + iterative EM-algorithm

It is worth tabulating the success rate of this algorithm at determining the degree of clustering or, rather, its ability to expose the number of probability densities responsible for the sample being analyzed. Again there were some very poor showings, see Table 1.10, and it appears that the more definitive the mean of the displaced cluster(s) the more likely this algorithm was going to result in a number of clusters greater than the number actually present. Indeed what is probably more striking is the extremely high rate of correct cluster *number* identification for the small sample size $n = 20$ in comparison with larger sample sizes, see Table 1.10. It was noticed, when applied to data sets of size $n = 20$ composed of 3 clusters, a high proportion of the simulations located 3 clusters, a main cluster and two outlying clusters $\epsilon = \epsilon_1 + \epsilon_2 = 0.4$, but a very low proportion of the Monte Carlo samples had the planted clusters exactly identified.

For this thesis, the above algorithm was then used in conjunction with an assessment of

the Silhouette value derived from an initial K-means estimate for cluster number. Instead of an initial K-means estimate enforced at 5, 4 or 3 clusters, for samples sizes of $n = 20,\ 50,\ 100$ respectively, we now observe the performance of the above algorithm based on an Expectation-Maximization algorithm with the initial estimate being that K-means estimate chosen by the best Silhouette value. So for example, a preliminary estimate of K-means cluster analysis was applied to each sample for various numbers of clusters and that number corresponding with the maximum Silhouette value was the chosen initial estimate for the EM-algorithm. The results for this scenario are contained in Table 1.11 and show that in comparison to the results in Table 1.9 that this EM-algorithm is heavily dependent upon the initial estimate and the added complexity may not really be necessary as it does not add much to the initial K-means estimate, optimized by Silhouette calculations. It is worth pointing to the success of this method for those clusters of outliers about a displaced mean of $d = 4\sqrt{\chi^2_{0.975}} = 10.8648$. The performance was excellent and even for solitary outliers removed to this distance the results were the best observed thus far for any algorithm.

n	ϵ	d	number of clusters	P_c
20	0		1	0.3461
	0.3	5.4324	2	0.4780
		10.8648	2	0.3653
	0.4	5.4324	3	0.9457
		10.8648	3	0.9777
50	0		1	0.6455
	0.3	5.4324	2	0.9527
		10.8648	2	0.6596
	0.4	5.4324	3	0.4752
		10.8648	3	0.3626
100	0		1	0.8153
	0.3	5.4324	2	0.8436
		10.8648	2	0.6048
	0.4	5.4324	3	0.4373
		10.8648	3	0.2517

Table 1.10: K-means + MINO + EM-algorithm: The success rate at determining cluster structure.

n	ϵ	d	p_t	$\bar{\alpha}$
20	0		0.5922	0.3797
	0.05	5.4324	0.7787	0.2923
		10.8648	0.9939	0.2643
	0.3	5.4324	0.9602	0.2966
		10.8648	> 0.9999	0.3000
50	0		0.2311	0.2237
	0.02	5.4324	0.3147	0.1493
		10.8648	0.721	0.0783
	0.3	5.4324	0.9564	0.3007
		10.8648	> 0.9999	0.3000
100	0		0.1374	0.1170
	0.01	5.4324	0.2065	0.1094
		10.8648	0.266	0.0998
	0.3	5.4324	0.9846	0.2998
		10.8648	> 0.9999	0.3000

Table 1.11: K-means + Silhouettes + MINO + iterative EM-algorithm.

Chapter 2

New Proposal

2.1 Univariate Adaptive Trimmed Likelihood

To expand on Clarke and Schubert (2006) and referring back to section 1.5 and the discussion on the LTS estimate for regression (1.6), this type of S-estimator may be arrived at from a more general form (Bednarski and Clarke 1993) where estimators known as trimmed likelihood estimators are defined. In the particular case of assuming a normal parametric density, and with approximately 50% trimming of the data, the trimmed likelihood estimator is equivalent to the LTS estimator. In Clarke (1994) an adaptive approach of estimating the proportion of trimming so as to minimize an estimate of the asymptotic variance of the estimator is applied. In this way the adaptive trimmed likelihood algorithm for *univariate* data has been applied as an adaptive regression estimator (Clarke 2000) to identify possible outliers, via the residuals, when observations are delineated by a linear regression.

We consider statistics by representing them as functionals (von Mises 1947) defined on the space of distribution functions ζ, where distributions are defined in the observation space $\Re^k$. The simplest situation is where dimension $p = 1$ whence, for example, the mean functional can be written $T[F] = \int x dF(x)$. Another example is the median functional

where $T[F] = F^{-1}(\frac{1}{2})$.

Naturally then it follows that the sample mean can be represented by

$$T[F_n] = \int x dF_n(x) = \overline{x}$$

where F_n is the empirical distribution function defined by the sample, for example,

$$F_n(y) = \frac{\#X_i's \leq y}{n}$$

and again the median functional

$$T[F_n] = F_n^{-1}(\frac{1}{2}) = \text{median}\{X_i\}_{i=1}^n.$$

The trimmed likelihood estimator was defined according to this functional approach by Bednarski and Clarke (1993). They introduced a trimmed likelihood principle where a proportion, α, of observations with the least probability of occurring, as deemed by the likelihood, are trimmed. This trimming is carried out in conjunction with simultaneously maximizing the likelihood, for example $T[F]$ is a solution to the estimating equation

$$L(F, \theta) = \int \phi(x, \theta) J[F\{y : \log f_\theta(y) \leq \log f_\theta(x)\}] dF(x) = 0$$

where $\phi(x, \theta) = \partial \log f_\theta(x)/\partial\theta$,

$$J(t) = \begin{cases} 0 & \text{if} \quad t \leq \alpha \\ 1 & \text{if} \quad \alpha < t \leq 1 \end{cases}$$

and f_θ is the parametric density associated with the form entertained for the common distribution of the $\{X_i\}$. Thus, for example, if $\alpha = 0$ so that no trimming is performed the estimator is defined as **the solution of the likelihood equation.** Visualize $T[F_n]$ is the solution of the equations

$$\int \phi(x, \theta) dF_n(x) = \frac{1}{n} \sum_{i=1}^n \phi(X_i, \theta) = 0.$$

If f_θ is chosen to be the normal location parametric family this simplifies to

$$\frac{1}{n}\sum(X_i - \mu) = 0 \Rightarrow \hat{\mu} = \overline{X} = T[F_n].$$

When the distribution F is symmetric with the centre of symmetry μ_0, then assuming a normal parametric density for $f_\theta(x)$, the resulting expansion for the statistic satisfies (Bednarski and Clarke 1993)

$$\sqrt{n}(T[F_n] - \mu_0) = \frac{\sqrt{n}\int_{\mu_0 - x_\alpha}^{\mu_0 + x_\alpha}(x - \mu_0)dF_n(x)}{1 - \alpha - 2x_\alpha f(x_\alpha)} + o_p(1).$$

From such an expansion one achieves

$$\sqrt{n}(T[F_n] - \mu_0) \xrightarrow{d} N(0, V(\alpha, F))$$

where the asymptotic variance equals

$$V(\alpha, F) = \frac{\int_{-x_\alpha}^{x_\alpha} x^2 dF_0}{\{1 - \alpha - 2x_\alpha f_0(x_\alpha)\}^2}. \tag{2.1}$$

Here $x_\alpha = F_0^{-1}(1 - \alpha/2)$ where F_0 is the underlying distribution F that is centred at zero and has density f_0.

The numerator in this expression is estimated by $(1 - \alpha)\overline{\sigma}_\alpha^2[F_n]$, where $\overline{\sigma}_\alpha^2[F_n]$ is defined (Clarke 1994) for all subsets of size $h = \lfloor \frac{n+1}{2} \rfloor, \ldots, n$ from an ordered sample $\{x_1, x_2, \ldots, x_n\}$ where $x_1 \leq x_2 \leq \ldots \leq x_n$, $\overline{x}^{(j)}$ is the average of $\{x_j, x_{j+1}, \ldots, x_{j+h-1}\}$ and

$$S^{(1)} = \frac{1}{h}\sum_{i=1}^{h}\{x_i - \overline{x}^{(1)}\}^2, \ldots, S^{(n-h+1)} = \frac{1}{h}\sum_{i=n-h+1}^{n}\{x_i - \overline{x}^{(n-h+1)}\}^2$$

which asymptotically behaves like (Bednarski and Clarke 1993)

$$\overline{\sigma}_\alpha^2[F_n] \approx \frac{1}{1 - \alpha}\int_{\mu_0 - x_\alpha}^{\mu_0 + x_\alpha}(x - \mu_0)^2 dF_n(x).$$

Thus the asymptotic normality result of the trimmed likelihood estimator for a fixed trimming proportion α, assuming a normal parametric family (Butler 1982, Rousseeuw 1983, 1984, Bednarski and Clarke 1993), ensures that (2.1), for $F_0(x) = F(x - \mu_0)$, becomes

$$V(\alpha, F_n) = \frac{(1 - \alpha)\overline{\sigma}_\alpha^2[F_n]}{\{1 - \alpha - \sqrt{\frac{2}{\pi}} z_{\alpha/2} e^{(-\frac{z_{\alpha/2}^2}{2})}\}^2}. \tag{2.2}$$

The adaptive trimmed likelihood algorithm (ATLA) is a procedure that estimates the $T_\alpha[F_n]$ which minimizes (2.2) for all α in the range of $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \lfloor\frac{n-1}{2}\rfloor\}$ (Clarke 1994, Bednarski and Clarke 2002) and *those observations that are trimmed, if any, in order to minimize (2.2) are considered outliers.*

2.2 Multivariate Adaptive Trimmed Likelihood

When analyzing multivariate normal data sets we are dealing with ellipsoidal probability densities of the form

$$\frac{1}{|\mathbf{\Sigma}|^{1/2}} f\left((\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma^{-1}} (\mathbf{x}-\boldsymbol{\mu})\right) \tag{2.3}$$

where $f : \Re^+ \to \Re^+$ is assumed to be nonincreasing, yielding a uni-modal density (Butler et al 1993), with $\boldsymbol{\mu}$ the centroid, $\mathbf{\Sigma}$ the covariance matrix and $|\mathbf{\Sigma}|$ indicating the determinant of $\mathbf{\Sigma}$ which is assumed to be non zero.

If the p-dimensional sample data is from a multivariate normal distribution with mean zero and covariance matrix being the identity, $N(\mathbf{0}, \boldsymbol{I}_p)$, then the sample covariance matrix using the MCD estimator of Butler et al. (1993) is such that

$$\hat{\mathbf{\Sigma}}_{1-\gamma}[F_n] \xrightarrow{wp1} \rho(\gamma)\boldsymbol{I}_p \tag{2.4}$$

where $\gamma = 1 - \alpha$ for an α proportion of trimming. Given the normalization required to ensure the integral of (2.3) equals 1, and transforming to polar co-ordinates (Davies 1987, Butler et al 1993), we arrive at the expression

$$\rho(\gamma)\boldsymbol{I}_p = \frac{1}{\gamma}\int_E \mathbf{x}\mathbf{x}^\top dF(\mathbf{x}) = \frac{2\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1} f(r^2) dr \boldsymbol{I}_p$$

where the first integral is over the set $E = \{\mathbf{x} | (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma^{-1}} (\mathbf{x}-\boldsymbol{\mu}) \leq r_\gamma^2\}$ for an r_γ^2 chosen so that $F\{E\} = \gamma$, where F represents the cumulative distribution function.

If in fact the data are multivariate normal, with known mean and covariance matrix, then it is well known that $r_\gamma^2 = \chi^2_{\gamma,p}$. Here $\chi^2_{\gamma,p}$ is the critical point of a chi squared distribution

with p degrees of freedom corresponding to the dimension of the data and having (γ) area under the chi squared density curve to the left of it. Γ is the usual gamma function, $\Gamma(v) = \int_0^\infty s^{v-1}e^{-s}ds$.

With regard to the sample estimate of the multivariate mean $\mathbf{T}[F_n]$, which in this context denotes the MCD estimate for location, we have

$$\sqrt{n}(\mathbf{T}[F_n] - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \kappa(\gamma)\boldsymbol{I}_p)$$

where (see Butler et al. 1993)

$$\kappa(\gamma) = \frac{p\Gamma(p/2)\int_0^{r_\gamma} r^{p+1}f(r^2)dr}{8\pi^{p/2}(\int_0^{r_\gamma} r^{p+1}f'(r^2)dr)^2} = \frac{\rho(\gamma)}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1}f'(r^2)dr)^2} \quad . \tag{2.5}$$

Here again $r_\gamma = \sqrt{\chi^2_{\gamma,p}}$ and $f(u) = (1/(2\pi))^{p/2}e^{-u/2}$ whence substituted in for f in (2.3) leads to the multivariate normal distribution.

If the data are generated from a multivariate normal the $\kappa(\gamma)$ above can be used to give an estimate for the asymptotic variance of $\mathbf{T}[F_n]$. Indeed if generated from a multivariate normal distribution $\hat{\Sigma}_{1-\gamma}[F_n]$ is asymptotically equal to $\rho(\gamma)\boldsymbol{I}_p$, (2.4), and for large n one need only locate a γ minimizing (2.5). Here we are concerned with potentially corrupt data and if the sample data consists of outliers then one would expect the value of $\rho(\gamma)\boldsymbol{I}_p$ to disagree with the sample variance $\hat{\boldsymbol{\Sigma}}_{1-\gamma}[F_n]$, which is no longer a covariance matrix from a normal distribution. With this in mind the multivariate extension of minimizing (2.1) becomes, as a direct consequence of (2.5), choose γ to minimize $|\frac{\kappa(\gamma)}{\rho(\gamma)}\hat{\boldsymbol{\Sigma}}_{1-\gamma}[F_n]|$ where:

$$|\frac{\kappa(\gamma)}{\rho(\gamma)}\hat{\boldsymbol{\Sigma}}_{1-\gamma}[F_n]| = \frac{|\hat{\boldsymbol{\Sigma}}_{1-\gamma}[F_n]|}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1}f'(r^2)dr)^{2p}} \tag{2.6}$$

In fact the above formula (2.6), which we will call the *Type 1 Proposal* (**T1**), is equivalent to minimizing $V(\alpha, F_n)$, for $\alpha = 1 - \gamma$, when $p = 1$ which is the preferred option 4 in Clarke (1994).

By Bednarski and Clarke (1993) for univariate data, $p = 1$ and F the cumulative standard

normal distribution, we see that

$$\gamma\hat{\Sigma}_{1-\gamma}[F_n] \xrightarrow{wp1} \int_{-z_{(1-\gamma)/2}}^{z_{(1-\gamma)/2}} x^2 dF(x).$$

This yields the Fisher consistent estimate for $\mathbf{\Sigma}$ which in the multivariate setting would result in choosing γ to minimize $|\frac{\kappa(\gamma)}{\rho(\gamma)}\gamma\hat{\mathbf{\Sigma}}_{1-\gamma}[F_n]|$, the *Type 2* (**T2**) *proposal*, where

$$|\frac{\kappa(\gamma)}{\rho(\gamma)}\gamma\hat{\mathbf{\Sigma}}_{1-\gamma}[F_n]| = \frac{|\gamma\hat{\mathbf{\Sigma}}_{1-\gamma}[F_n]|}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1}f'(r^2)dr)^{2p}}\ . \tag{2.7}$$

For $p = 1$ minimizing the objective function (2.7) is equivalent to choosing option 5 in Clarke (1994). We explore the comparison between the **T1** and **T2** proposals using Monte Carlo simulations.

These objective functions, (2.6) and (2.7), can help us to identify outlying data since there is, by Theorem 2 in Butler et al (1993), a specific subset of a sample data set which will minimize them. That subset corresponding to this minimum asymptotic variance, of a robust estimate for location, will be considered *free* of outliers. Such a decision is based on the Fisher Information argument that any reduction in the *information* a sample contains necessarily increases the variance of any estimate for any parameter. When a data set is trimmed of data, the necessary reduction in information should therefore increase the variance of an estimate for location. This should always be the case *unless* the data being removed from the sample is contamination.

2.3 Basic constructs for new algorithm

The first step in the new proposal involves obtaining a robust MCD estimate for the centroid and scale of the sample data set followed by a Forward Search for subsets which minimize the objective function. The procedure, therefore, is to initially arrange each sample point in ascending order of Mahalanobis distances, M_i, from this MCD estimate for the centroid $\hat{\boldsymbol{\mu}}$,

$$M_i = \sqrt{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})}, \tag{2.8}$$

where $\hat{\boldsymbol{\Sigma}}$ is the covariance of those observations contributing to the MCD estimate.
The data is ordered, then divided into two subsets, first the set of h points closest to the centroid, that is those points responsible for the MCD estimate, and a set of $n-h$ points being tested for outlyingness. The data is arranged thus,

$$\{\mathbf{x}_{i_{M_{(1)}}}, \mathbf{x}_{i_{M_{(2)}}}, \ldots, \mathbf{x}_{i_{M_{(h)}}}, \mathbf{x}_{i_{M_{(h+1)}}}, \ldots, \mathbf{x}_{i_{M_{(n)}}}\}$$

where $h = \lfloor \frac{n+p+1}{2} \rfloor$ for p-dimensional data and $M_{(1)} \leq M_{(2)} \leq \ldots \leq M_{(n)}$.

Once the objective function, either (2.6) or (2.7), has been calculated using the MCD subset, the subset is inflated to include the nearest point, that is, that observation in the complement of the subset closest to the subset, whence the whole sample is again arranged in ascending order of Mahalanobis distance using the centroid and covariance matrix derived from this inflated subset. With this step, and each subsequent repetition thereof, it is imperative to note that some members of the *previously* assessed subset can interchange with complement members *before* a further inflation due to the re-ordering. (Atkinson, Riani and Cerioli 2004). The objective function being used is again exerted on this new subset before the subset is again incremented to include the nearest observation. This procedure is repeated until the subset has been inflated to include the whole sample set. Ideally that subset yielding the minimum for the objective function being examined will be regarded as outlier-free, so observations, if any, not a member of this subset are identified as outlying.

In the coming sections we see that it is more appropriate to choose any minimum, local or global, occurring for an $\alpha > 0$ and if no such minimum exists, then the data set can be considered outlier free. The coming sections also pose the enigma of sometimes having to choose between multiple minima for $\alpha > 0$. Corresponding to each minimum

$$\mathsf{m}_i$$

where $i = 1, \ldots, j$ for an increasing $\alpha > 0$ we have a subset

$$S_{\gamma_{\mathsf{m}_i}}$$

of retained data. Do we choose the subset corresponding to the minimum of all minima m_i occurring for an $\alpha > 0$,

$$S_{\gamma_{\min_i(\mathsf{m}_i)}},$$

or do we choose the minimum subset of retained data, which will correspond with the greatest α,

$$(S_{\gamma_{\mathsf{m}_j}})?$$

We shall see that in most cases these two will agree.

2.4 Monte Carlo simulations

We conducted Monte Carlo simulations for both $p = 2$ and $p = 4$ dimensional data sets to assess the efficacy of the Type 1 (**T1**) and Type 2 (**T2**) proposals. Data sets distributed normally, then contaminated with a pre-specified proportion, ϵ, of outliers were subjected to the new proposal with the proportion of outliers detected and trimming frequency recorded. Sample sizes of $n = 20,\ 50,\ 100,\ 500$ were generated with a proportion $\epsilon = 0,\ 1/n,\ 0.1,\ 0.3,\ 0.4$ of the data shifted from the main mean.

It was discovered that **T2** was too sensitive when applied to small data sets. **T2** was, for instance, identifying more than 20% of *clean* bivariate data sets of size $n = 20$ as outliers. Table 2.1 contains the simulation results for 2 and 4 dimensional, clean data sets of sizes $n = 10, 15, 20, 25, 30, 40, 50$ and the proportion of samples where at least one observation was identified as outlying, p_t, and the average number of observations identified as outlying, $\bar{\alpha}$. Ideally our p_t and $\bar{\alpha}$ should both be zero when applying outlier detection methods to clean data sets and it also noted that p_t and $\bar{\alpha}$ are sample statistics and do not necessarily measure the exact *planted* outlier detection proportions. For samples of size $n = 30$ Table 2.1 shows the average proportion of observations identified as outlying by **T2** is 0.0249 for bivariate data and 0.0366 for 4 dimensional samples. As the sample

size increases it becomes increasingly less likely that any normally distributed observations will be identified as outliers.

Given these results it was decided to use **T1** for data sets of size $n < 30$ and to apply **T2** otherwise since **T1** was less sensitive than **T2**.

n	p	**T1** p_t	$\bar{\alpha}$	**T2** p_t	$\bar{\alpha}$
10	2	0.229	0.0766	0.423	0.1412
15		0.095	0.0307	0.221	0.0670
20		0.063	0.0208	0.211	0.0689
25		0.025	0.0079	0.111	0.0337
30		0.011	0.0027	0.076	0.0249
40		0.002	0.0001	0.025	0.0058
50		< 0.001	< 0.0001	0.007	0.0029
10	4	0.446	0.1231	0.766	0.2120
15		0.128	0.0365	0.426	0.1179
20		0.057	0.0193	0.319	0.1031
25		0.026	0.0084	0.156	0.0506
30		0.008	0.0032	0.111	0.0366
40		0.001	0.0003	0.02	0.0081
50		0.001	0.0002	0.003	0.0012

Table 2.1: Establishing **T2** cut-off sample size.

Now returning to the simulations regarding contaminated data, with $\epsilon = 1/n,\ 0.1,\ 0.3$, this outlying proportion of data consisted of a pth variable centred about a displacement d from the *clean* data distributed $N([0,0]^\top, \boldsymbol{I_2})$, $N([0,0,0,0]^\top, \boldsymbol{I_4})$ respectively, where $d = q\sqrt{\chi^2_{0.975,2}},\ q\sqrt{\chi^2_{0.975,4}}$ for $q = 2,\ 4$ (Juan and Prieto 2001):

- For $q = 2$ we have $N([0, 5.4324]^\top, \boldsymbol{I_2})$ and $N([0,0,0,6.6763]^\top, \boldsymbol{I_4})$.
- For $q = 4$ we have $N([0, 10.8348]^\top, \boldsymbol{I_2})$ and $N([0,0,0,13.3526]^\top, \boldsymbol{I_4})$.

When assessing the data sets with a proportion $\epsilon = 0.4$ of contamination, these outliers formed two clusters of outlying data, each cluster representing $0.2n$ of the whole data set. These two clusters consisted of a proportion, $\epsilon_1 = \epsilon_2 = 0.2$, of the pth variable centred about a displacement of $\pm d$ respectively.

Table 2.2 contains the results when **T1** was applied to data sets of size $n = 20$ and **T2** was applied to samples of size $n = 50, 100, 500$, each sample contaminated by a single outlier $\epsilon = 1/n$. The Monte Carlo samples assessed were, respectively, $p = 2$ and $p = 4$ dimensional and Table 2.2 shows us that with regard to the larger outlier displacement mean of the pth

Table 2.2: Simulation results for sole outlier.

dimension p=2					*dimension p=4*				
n	ϵ	d	p_t	$\bar{\alpha}$	n	ϵ	d	p_t	$\bar{\alpha}$
20	0.05	5.4324	0.66	0.0386	20	0.05	6.6763	0.479	0.0374
		10.8648	> 0.999	0.0513			13.3526	> 0.999	0.0506
50	0.02	5.4324	0.716	0.0158	50	0.02	6.6763	0.775	0.0161
		10.8648	> 0.999	0.0207			13.3526	> 0.999	0.0202
100	0.01	5.4324	0.705	0.0072	100	0.01	6.6763	0.81	0.0101
		10.8648	> 0.999	0.0101			13.3526	> 0.999	0.0102
500	0.002	5.4324	0.509	0.001	500	0.002	6.6763	0.732	0.0015
		10.8648	> 0.999	0.002			13.3526	> 0.999	0.002

variable, for example, of sample size $n = 500$, $p = 4$ and $d = 4\sqrt{\chi^2_{0.975,4}}$, the solitary outlier was always identified. For the smaller outlier displacement $d = 2\sqrt{\chi^2_{0.975,4}}$, $p = 4,\ n = 500$, Table 2.2 shows this outlier was identified in more than 70% of samples.

Table 2.3 shows the results when the new proposal, **T1** and **T2** accordingly, were applied to samples contaminated with one outlying cluster. The proportion of data belonging to this cluster shifted about about a mean displacement $d = q\sqrt{\chi^2_{0.975,p}}$ was $\epsilon = 0.1$ and $\epsilon = 0.3$ respectively.

For large sample sizes, $n = 500$, the results were excellent for both displacement means. Even for samples as small as $n = 50$ this algorithm is detecting the outlying cluster in more than 90% of samples with the cluster centred about the smaller displacement and nearly always for the larger displacement.

The new proposal's ability to detect point mass outliers was also assessed using Monte Carlo samples with one outlying cluster. In these samples the cluster proportions, $\epsilon = 0.1$ and $\epsilon = 0.3$, were distributed $N([0, q\sqrt{\chi^2_{0.975,2}}]^\top, 0.1\boldsymbol{I_2})$ and $N([0,0,0, q\sqrt{\chi^2_{0.975,4}}]^\top, 0.1\boldsymbol{I_4})$. Table 2.4 contains excellent results for samples of size $n \geq 50$. The results show that for nearly every sample the point mass outlying cluster was detected. If one was to *trim* the data set of these observations identified as outliers, it may be the case of a very slight tendency to overtrim, for example $p = 4$, $n = 50$, $\epsilon = 0.3$ and $d = 4\sqrt{\chi^2_{0.975,4}}$ we see $\bar{\alpha} = 0.3113$. Of course this is only a slight loss in efficiency whilst the parameter estimates

Table 2.3: Simulation results one outlying cluster.

dimension p=2					*dimension p=4*				
n	ϵ	d	p_t	$\bar{\alpha}$	n	ϵ	d	p_t	$\bar{\alpha}$
20	0.1	5.4324	0.652	0.0721	20	0.1	6.6763	0.600	0.0712
		10.8648	> 0.999	0.1025			13.3526	> 0.999	0.1017
	0.3	5.4324	0.641	0.2001		0.3	6.6763	0.542	0.1699
		10.8648	> 0.999	0.3083			13.3526	0.992	0.3027
50	0.1	5.4324	0.926	0.0912	50	0.1	6.6763	0.953	0.0959
		10.8648	> 0.999	0.1031			13.3526	> 0.999	0.1016
	0.3	5.4324	0.912	0.2799		0.3	6.6763	0.925	0.283
		10.8648	> 0.999	0.3104			13.3526	> 0.999	0.3058
100	0.1	5.4324	0.973	0.0952	100	0.1	6.6763	0.987	0.0985
		10.8648	> 0.999	0.1026			13.3526	> 0.999	0.1012
	0.3	5.4324	0.958	0.2912		0.3	6.6763	0.987	0.2997
		10.8648	> 0.999	0.3077			13.3526	> 0.999	0.3042
500	0.1	5.4324	> 0.999	0.0992	500	0.1	6.6763	> 0.999	0.1003
		10.8648	> 0.999	0.1022			13.3526	> 0.999	0.1009
	0.3	5.4324	> 0.999	0.3032		0.3	6.6763	> 0.999	0.3004
		10.8648	> 0.999	0.3061			13.3526	> 0.999	0.3008

will not be significantly impacted. The problem of trimming data only becomes a serious issue if data sets are undertrimmed, that is some outliers are *not* detected or when data sets are severely overtrimmed, resulting in the loss of too much clean data which will greatly reduce estimate efficiency.

Table 2.4: Simulation results for one cluster of Point Mass outliers.

dimension p=2					*dimension p=4*				
n	ϵ	d	p_t	$\bar{\alpha}$	n	ϵ	d	p_t	$\bar{\alpha}$
20	0.1	5.4324	0.967	0.1556	20	0.1	6.6763	0.939	0.1819
		10.8648	> 0.999	0.1566			13.3526	> 0.999	0.1887
	0.3	5.4324	0.961	0.3213		0.3	6.6763	0.878	0.2948
		10.8648	> 0.999	0.3337			13.3526	> 0.999	0.3318
50	0.1	5.4324	0.999	0.1056	50	0.1	6.6763	0.999	0.1029
		10.8648	> 0.999	0.1084			13.3526	0.989	0.1039
	0.3	5.4324	0.998	0.3153		0.3	6.6763	> 0.999	0.3068
		10.8648	> 0.999	0.3158			13.3526	> 0.999	0.3113
100	0.1	5.4324	> 0.999	0.1028	100	0.1	6.6763	> 0.999	0.1013
		10.8648	> 0.999	0.1028			13.3526	> 0.999	0.1012
	0.3	5.4324	> 0.999	0.3093		0.3	6.6763	> 0.999	0.3045
		10.8648	> 0.999	0.3097			13.3526	> 0.999	0.3049
500	0.1	5.4324	> 0.999	0.1023	500	0.1	6.6763	> 0.999	0.1010
		10.8648	> 0.999	0.1021			13.3526	> 0.999	0.1010
	0.3	5.4324	> 0.999	0.3065		0.3	6.6763	> 0.999	0.3033
		10.8648	> 0.999	0.3063			13.3526	> 0.999	0.3030

Table 2.5: Simulation results for two outlying clusters.

dimension p=2					*dimension p=4*				
n	ϵ	d	p_t	$\bar{\alpha}$	n	ϵ	d	p_t	$\bar{\alpha}$
20	0.4	5.4324	0.587	0.2384	20	0.4	6.6763	0.372	0.1438
		10.8648	> 0.999	0.406			13.3526	0.941	0.3752
50	0.4	5.4324	0.909	0.3713	50	0.4	6.6763	0.904	0.3692
		10.8648	> 0.999	0.4116			13.3526	> 0.999	0.4089
100	0.4	5.4324	0.963	0.3889	100	0.4	6.6763	0.979	0.3961
		10.8648	> 0.999	0.4083			13.3526	> 0.999	0.4056
500	0.4	5.4324	0.999	0.4039	500	0.4	6.6763	> 0.999	0.4035
		10.8648	> 0.999	0.4072			13.3526	> 0.999	0.4039

Table 2.5 contains the results pertaining to the new proposals application to samples contaminated by *two* outlying clusters, each of size $0.2n$ leading to an overall proportion of $\epsilon = 0.4$ of the data set is outlying. The figures show the same level of success as when applied to data sets with one outlying cluster.

2.4.1 Instances of multiple minima

The Monte Carlo series discussed above all show a very strong success rate at identifying outlying observations and *not* identifying clean observations as outlying as the sample size increases. For small samples $n = 20$ contaminated by shifted observations about the smaller displacement the results are not as strong, but many of these contaminants may not have been *technically* outlying. When outliers are distributed about a small displacement from the main population, a certain percentage may not be outlying.

With regard to sample data sets contaminated with a solitary outlier, it is important to note that a *global* minimum occurred for an $\alpha = 1/n$ on more than 99% of occasions. Clustered contamination rarely forced a *global* minimum away from $\alpha = 0$, but for the high proportion of correctly identified outlying clusters there occurred a *local* minimum for an $\alpha > 0$.

When confronted with data sets contaminated by outlying clusters we found increasing instances of multiple minima for $\alpha > 0$. The larger the proportion of clustered outliers, the more frequent the occurrence of multiple minima. Table 2.6 contains the proportion of such instances, p_M, for an $\epsilon = 0.1,\ 0.3$, for samples of size $n = 50,\ 100$ in both $p = 2$ and $p = 4$ dimensional cases.

If there occurs a global minimum for some $\alpha > 0$ this will be taken as the trimming proportion necessary, otherwise, we consider the proportion of trimming required by the data to be governed by any local minimum away from $\alpha = 0$. The estimated chance of more than 1 local minimum away from $\alpha = 0$ is tabulated in Table 2.6 for various levels of contaminant data for any sample size n.

When using **T2**, multiple minima occur more frequently for small samples and those contaminated with clustered outliers. **T1** is less sensitive than **T2** and so, as expected, yielded less chance of multiple minima. When existence of multiple minima occurred it was noticed in the majority of instances that the subset of retained data, associated with the minimum corresponding to the greatest α, coincided with the subset associated with the *minimum* value of the minima occurring,

$$S_{\gamma_{\min_i(\mathtt{m}_i)}} = S_{\gamma_{\mathtt{m}_j}}.$$

The **T2** proposal was applied to bivariate data sets of size $n = 100$ with proportions $\epsilon = 0.1,\ 0.3$ of each generated sample corrupted with shifted means of $d = 2\sqrt{\chi^2_{0.975,2}}$ and $d = 4\sqrt{\chi^2_{0.975,2}}$ respectively. Collected, for Figures 2.1-2.6, are instances of multiple minima occurring for $\alpha > 0$ and the relationship between the size of the retained subset and the corresponding value of the objective function (2.7) are illustrated.

For $d = 4\sqrt{\chi^2_{0.975,2}}$, Figures 2.1 and 2.2 depict a steep fall in the objective function (2.7), in the vicinity of the correct trimming proportion, noting these plots represent cases of multiple minima for $\alpha > 0$ *only*. For the majority of cases only one minima occurs for $\alpha > 0$ and the plots (not shown) for these cases depict the same dramatic fall in the value of (2.7).

Table 2.6: Frequency of multiple minima.

dimension p=2				*dimension p=4*			
n	ϵ	d	p_M	n	ϵ	d	p_M
20	0		0.014	20	0		0.002
	0.02	5.4324	0.050		0.02	6.6763	0.040
		10.8648	0.110			13.3526	0.059
	0.1	5.4324	0.123		0.1	6.6763	0.074
		10.8648	0.078			13.3526	0.094
	0.3	5.4324	0.151		0.3	6.6763	0.122
		10.8448	0.199			13.3526	0.141
50	0		0	50	0		0
	0.02	5.4324	0.003		0.02	6.6763	0.004
		10.8648	0.008			13.3526	0.004
	0.1	5.4324	0.091		0.1	6.6763	0.035
		10.8648	0.042			13.3526	0.045
	0.3	5.4324	0.153		0.3	6.6763	0.087
		10.8448	0.101			13.3526	0.073
100	0		0	100	0		0
	0.01	5.4324	< 0.001		0.01	6.6763	< 0.001
		10.8648	< 0.001			13.3526	< 0.001
	0.1	5.4324	0.17		0.1	6.6763	0.069
		10.8648	0.023			13.3526	0.013
	0.3	5.4324	0.168		0.3	6.6763	0.043
		10.8448	0.066			13.3526	0.024

Closer inspections, on reduced intervals of subset size for clarity with $d = 4\sqrt{\chi^2_{0.975,2}}$, see Figures 2.3 and 2.4, expose the multiple minima and we notice the *minimum*, $\min_i(\mathsf{m}_i)$, of these correspond with the greatest α for which a minimum occurs, $\min_i(\mathsf{m}_i) = \mathsf{m}_j$. Figures 2.5-2.6 display examples of multiple minima for the smaller shift in outlier mean, $d = 2\sqrt{\chi^2_{0.975,2}}$, where the drop in the value of (2.7) was less obvious. Again we can see the minimum corresponding to the greatest α coincided with the minimum value of the minima occurring.

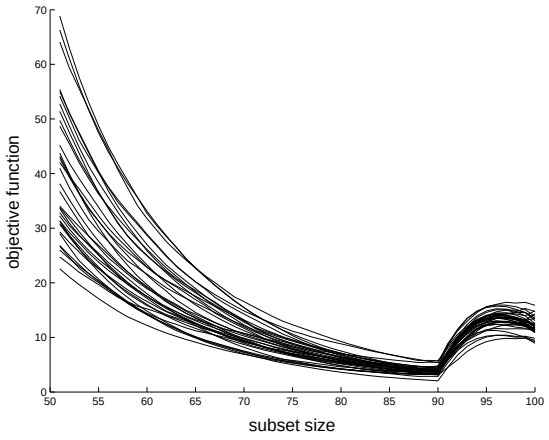


Figure 2.1: $n = 100$, $\epsilon = 0.1$, $d = 4\sqrt{\chi^2_{0.975,2}}$.

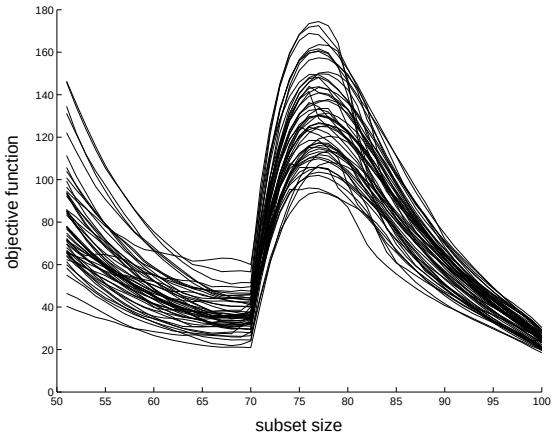


Figure 2.2: $n = 100$, $\epsilon = 0.3$, $d = 4\sqrt{\chi^2_{0.975,2}}$.

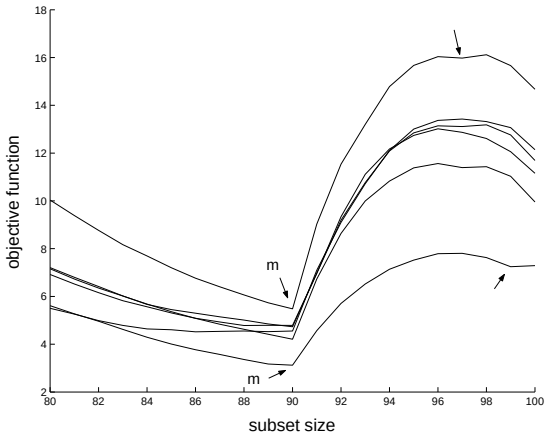


Figure 2.3: $n = 100$, $\epsilon = 0.1$, $d = 4\sqrt{\chi^2_{0.975,2}}$.

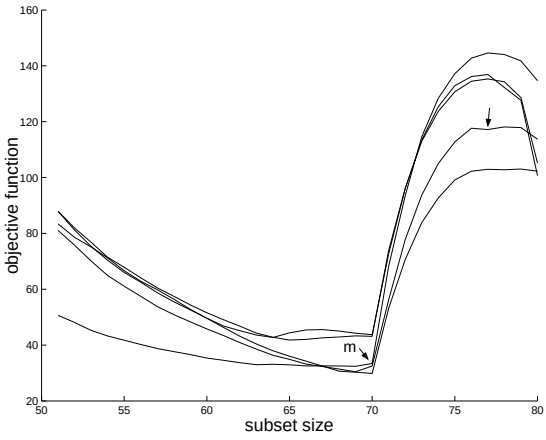


Figure 2.4: $n = 100$, $\epsilon = 0.3$, $d = 4\sqrt{\chi^2_{0.975,2}}$.

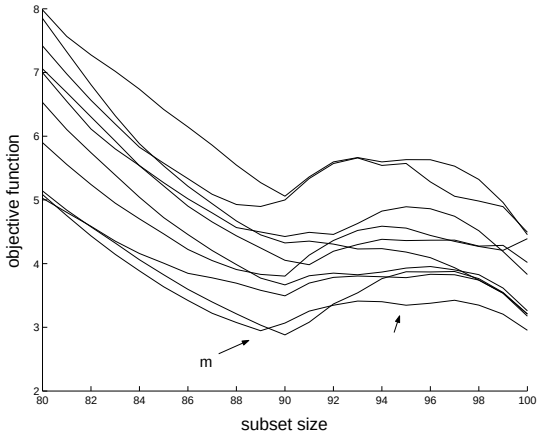


Figure 2.5: $n = 100$, $\epsilon = 0.1$, $d = 2\sqrt{\chi^2_{0.975,2}}$.

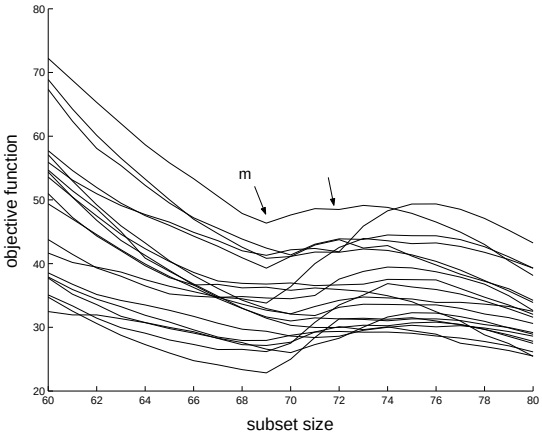


Figure 2.6: $n = 100$, $\epsilon = 0.3$, $d = 2\sqrt{\chi^2_{0.975,2}}$.

2.4.2 t-distributed data

The quintessential argument of this thesis is that the proposals, **T1** and **T2**, assess *normality*. If the application of **T1** or **T2** results in *no* minima for an $\alpha > 0$ then the sample is assumed to be *normal*. Any minima occurring for $\alpha > 0$ are assumed to coincide with a proportion of *non*-normal data. Thus when data sets are not governed by the normal probability density, 'heavy tails' may be present and outliers should be detected if one supposes the data sets are normally distributed. In this section we observe the impact of the **T2** proposal on samples governed by t-distributions, measuring the asymptotic proportion, α, of data identified as outlying if normality is assumed.

Returning to (2.5)

$$\kappa(\gamma) = \frac{p\Gamma(p/2)\int_0^{r_\gamma} r^{p+1} f(r^2)dx}{8\pi^{p/2}(\int_0^{r_\gamma} r^{p+1} f'(r^2)dr)^2} = \frac{\rho(\gamma)}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{y} r^{p+1} f'(r^2)dr)^2} \tag{2.9}$$

Finding the subset minimizing $\kappa(\gamma)$ in (2.9) will correspond with a particular value of $\alpha = (1-\gamma)$, or proportion of trimming necessary to arrive at a data set free of outliers given the ρ function in the numerator is now consistent with the probability density governing the t-distribution. The multivariate t-distribution has a probability density function of the form (Mardia, Kent and Bibby 1979)

$$f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{c_p |\boldsymbol{\Sigma}|^{-1/2}}{[1 + \frac{1}{\upsilon}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]^{(\upsilon+p)/2}}$$

where

$$c_p = \frac{\Gamma(\frac{\upsilon+p}{2})}{(\upsilon\pi)^{p/2}\Gamma(\frac{\upsilon}{2})}$$

and υ is the degrees of freedom.

When data assumed normal is in fact t-distributed, $\hat{\boldsymbol{\Sigma}}_\alpha[F_n]$ in equation (2.6) asymptotically goes to $\rho_{t_\upsilon}\boldsymbol{I}_p$ so the asymptotic minimum of the objective function (2.9) can be seen as the *global* minimum of

$$\nu(\alpha, F_n) = \left(\rho_{t_\upsilon}\frac{\kappa(\gamma)}{\rho(\gamma)}\right)^p = \left| \frac{\frac{2\pi^{p/2}}{p\Gamma(p/2)}\int_0^{y} \mathbf{x}^{p+1}\frac{\Gamma(\frac{\upsilon+p}{2})}{(\upsilon\pi)^{p/2}\Gamma(\frac{\upsilon}{2})}\frac{1}{(1+\frac{1}{\upsilon}\mathbf{x}^2)^{((\upsilon+p)/2)}}d\mathbf{x}}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{y}\mathbf{x}^{p+1}f'(\mathbf{x}^2)d\mathbf{x})^2} \right| \tag{2.10}$$

where the limit y here corresponds to $t_{(1-\alpha),\upsilon}$.

Figures 2.7-2.11 depict the proportions, $\gamma = (1 - \alpha)$, of t-distributed data that will be retained given normality of the data has been assumed as $n \to \infty$.

Figures 2.7-2.8 depict the asymptotic minimum of (2.10) for t-distributed data with 1 degree of freedom which is equivalent to data sets distributed according to the Cauchy probability density.

Figure 2.7 illustrates when the asymptotic minimum is reached for bivariate Cauchy distributed data. It is interesting to note that the increase in dimension here has increased the amount of trimming necessary to minimize (2.10). In the univariate case, not depicted, the minimum occurred for $\alpha \approx 0.15$ but in the bivariate case we have the global minimum occurring at $\alpha \approx 0.25$. Thus the 25% of the data with the greatest Mahalanobis distance from the estimate for centroid is identified as outlying from the normal perspective. Figure 2.8 is consistent with an increase in dimension, $p = 3$, resulting in an increase in the proportion, $\alpha \approx 0.30$, of outliers detected and needing deletion to minimize (2.10).

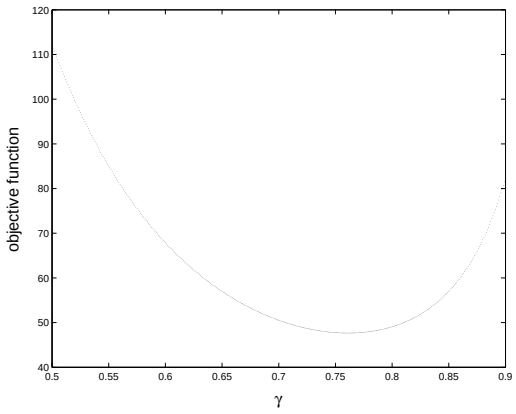


Figure 2.7: Bivariate Cauchy.

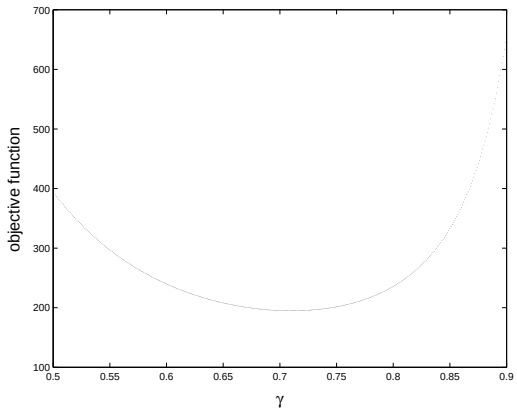


Figure 2.8: Trivariate Cauchy.

Table 2.7 contains the average trimming proportions, $\bar{\alpha}$, when the **T2** proposal is applied to bivariate and trivariate Cauchy distributed data. Monte Carlo samples of size $n = 20, 30, 40, 50, 100,$ respectively, were examined to establish a *cut off* sample size for when **T2** is expected to yield a $\bar{\alpha}$ consistent with the asymptotic minimums shown in Figures

2.7-2.8. The corresponding measure of γ would indicate when **T2** is starting to perform efficiently. It can be seen that the asymptotic minimum is reached for sample sizes as low as $n = 20$ which is a somewhat counterintuitive result given the over sensitivity of the new proposal when dealing with normally distributed samples of $n = 20$. This is better than expected result, the more *abnormal* a data set the more sensitive **T2** is to outliers, for small as well as large samples if the data is considered normally distributed.

Figures 2.9-2.10 depict cases investigated for t-distributed data sets with $\upsilon = 3$ degrees of freedom. The proportion of outliers requiring trimming to minimize (2.10) again increases in concert with an increase in dimension, but it is important to recognize, that the associated increase in the number of degrees of freedom explains a definitive reduction in the proportion of outliers detected. This is to be expected as the t-distribution converges in distribution to normality, whence there should be *no* outliers detected.

In the case of bivariate t_3-distributed data we can see in Figure 2.9 that the subset of observations responsible for the greatest 4% of Mahalanobis distances from the parameter value for location will be identified as outliers if one models the data as normally distributed. Figure 2.10 shows around 5% of trivariate t_3-distributed data is identified as outlying.

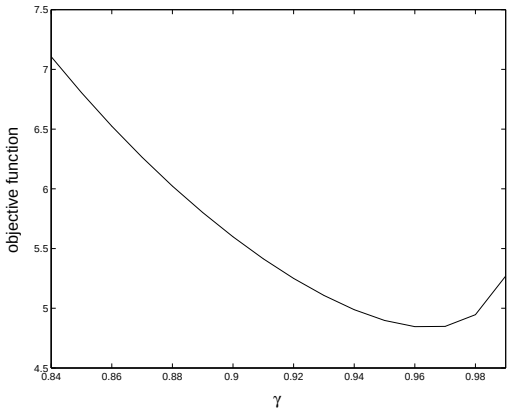


Figure 2.9: Bivariate t_3-distributed data.

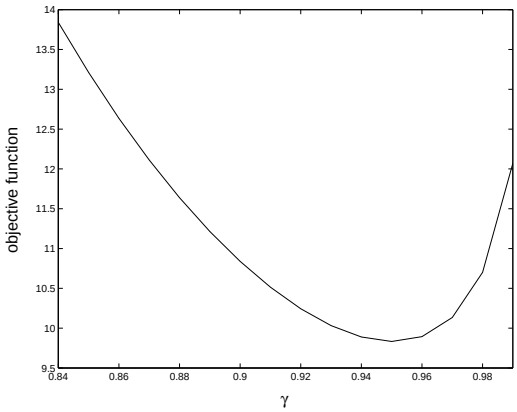


Figure 2.10: Trivariate t_3-distributed data.

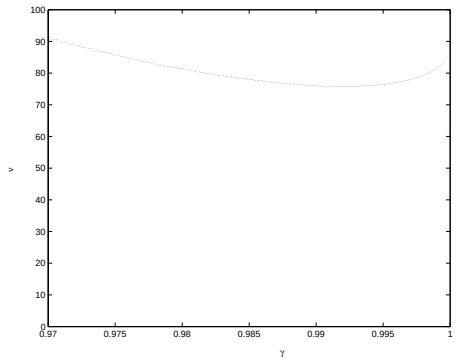


Figure 2.11: $p = 20$ dimensional, t_{10}-distributed data.

Table 2.8 shows us that the asymptotic minimum is reached for samples of size $n \approx 50$ if one is dealing with t_3-distributed data sets. Of course the analyst may not know if the random data set before them is t_3-distributed or not. If the data set is of size ≥ 50 then we can be sure this proposal **T2** will successfully detect the outliers if so.

n	p	$\bar{\alpha}$
20	2	0.2372
	3	0.2602
30	2	0.2303
	3	0.2729
40	2	0.2288
	3	0.2750
50	2	0.2264
	3	0.2853
100	2	0.2306
	3	0.2886

Table 2.7: t_1 data.

n	p	$\bar{\alpha}$
20	2	0.1014
	3	0.1150
30	2	0.0681
	3	0.0805
40	2	0.0539
	3	0.0600
50	2	0.0407
	3	0.0545
100	2	0.0376
	3	0.0502
120	2	0.0363
	3	0.0497

Table 2.8: t_3 data.

p	n	$\bar{\alpha}$
20	40	0.1218
	50	0.0328
	60	0.0124
	70	0.0093
	80	0.0090
	90	0.0071
	100	0.0079
	120	0.0075

Table 2.9: t_{10} data.

Tests were also conducted for t-distributed data with 10 degrees of freedom and we must recall that normally distributed data is ideally free of outliers. It is of no surprise, therefore, to find that for t_{10}-distributed data of $p = 1, 2, 3$ dimensions, the asymptotic global minimum occurred at $\alpha = 0$. Increasing dimension has been shown to increase the proportion of observations identified as outliers in t-distributed data. With this in mind we illustrate the size of $\gamma = (1 - \alpha)$ when **T2** is applied to twenty dimensional, $p = 20$, t-distributed data with $\upsilon = 10$ degrees of freedom. On average, if normality was assumed, between 0.5% and 1% of data with greatest Mahalanobis distance from the estimate for location will be identified as suspiciously outlying.

Table 2.9 shows us that for 20 dimensional t_{10}-distributed data, one can expect **T2** not to overtrim when sample sizes of $n \geq 50$ are being assessed.

2.4.3 Correlated transformations

For further confirmation of the versatility of the **T1**, **T2** proposals, their consistency when applied to samples of a definitive shape was investigated. For this algorithm to be reliable it must perform independently of how a data set may be correlated.

Suppose we generate a multivariate normal data set

$$\mathbf{Z} = (Z_1, \ldots, Z_p)^\top \sim N_p(\mathbf{0}, \boldsymbol{I}_p).$$

The correlation matrix, $\boldsymbol{P}$, for a vector random variable, $\mathbf{W} = (W_1, \ldots, W_p)^\top$, having non-singular covariance matrix, $\boldsymbol{\Sigma}$, can be represented as (Chatfield and Collins 1980),

$$\mathbf{P} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^{-1} \tag{2.11}$$

where

$$\boldsymbol{\Lambda} = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \sigma_p \end{pmatrix}$$

for $\sigma_i^2 = \mathrm{Var}(W_i)$ consistent with the sample covariance $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{W})$. If we substitute for $\boldsymbol{P}$ a pre-specified $\tilde{\boldsymbol{P}}$ and re-arrange (2.11) we arrive at

$$\boldsymbol{\Lambda}\tilde{\boldsymbol{P}}\boldsymbol{\Lambda} = \boldsymbol{\Sigma}$$

such that for a $p \times p$ nonsingular matrix $\mathbf{A}$,

$$\mathbf{A} = (\boldsymbol{\Lambda}\tilde{\boldsymbol{P}}\boldsymbol{\Lambda})^{1/2} \Rightarrow \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top.$$

Therefore putting

$$\mathbf{X} = \mathbf{A}\mathbf{Z},$$

where $\mathbf{X} = (X_1, \ldots, X_p)$, results in the covariance

$$\text{cov}(\mathbf{X}) = \text{cov}(\mathbf{AZ}) = \mathbf{A}\text{cov}(\mathbf{Z})\mathbf{A}^T = \mathbf{AA}^T = \mathbf{\Sigma}.$$

Using this fact we generated Monte Carlo simulations for bivariate data sets $\mathbf{X}$ with a pre-specified correlation

$$\tilde{\boldsymbol{P}} = \begin{pmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{pmatrix}$$

for $\rho_{12} = \rho_{21} \approx -0.95, -0.50, 0, +0.50, +0.95$ respectively.

Figures 2.12-2.16 depict the *shape* of the simulations corresponding with the pre-specified correlations. The proportion of samples, p_t, for which outliers were identified by **T1** and **T2**, depending on the sample size, and the average proportion identified per sample, $\bar{\alpha}$, was recorded in Table 2.10 for samples of size $n = 20, 50, 100$.

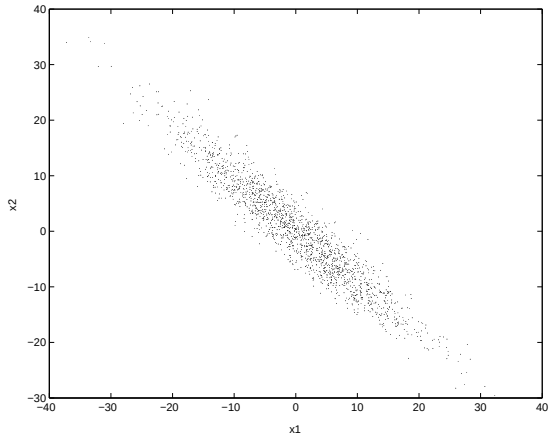


Figure 2.12: $\rho_{12} = \rho_{21} \approx -0.95$.

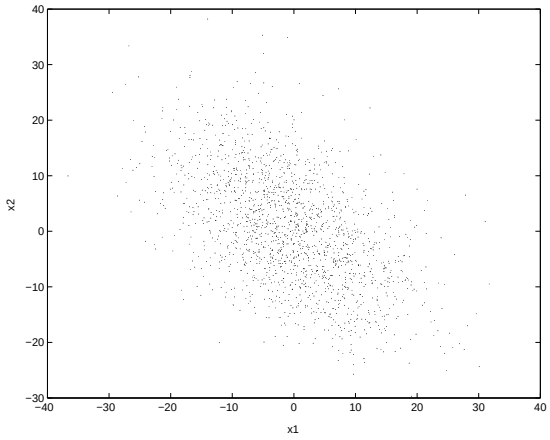


Figure 2.13: $\rho_{12} = \rho_{21} \approx -0.50$.

For each of the sample sizes $n = 20$, $n = 50$ and $n = 100$ tested for sample types shown in Figures 2.12-2.16, the magnitude of the figures is not the issue, rather their consistency. A simple application of a chi-squared test, 1% significance level, predicate that the results of both the **T1** and **T2** proposals are independent of the shape of the data being analyzed.

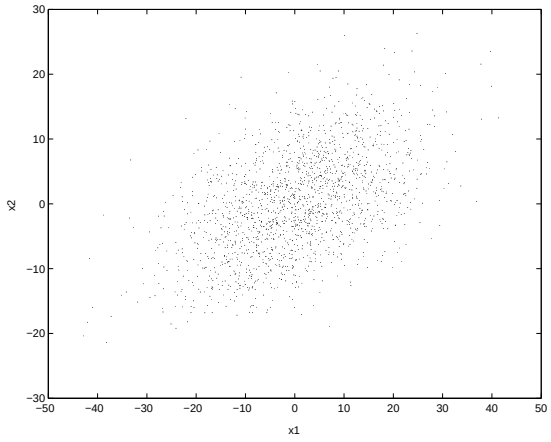


Figure 2.14: $\rho_{12} = \rho_{21} \approx +0.50$.

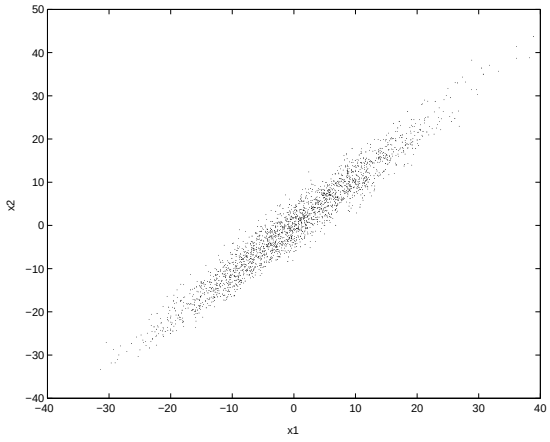


Figure 2.15: $\rho_{12} = \rho_{21} \approx +0.95$.

2.4.4 T2 vs non-robust estimates

Figure 2.17 depicts the impact of not removing a single, planted outlier from a sample of size $n = 100$ and dimension $p = 3$ as the outlier mean increases over the range $d = 0, \ldots, 20$. For simulations of size $N = 1000$ for *each* of the displaced means the median MSE obtained when the single outlier is not removed is compared with the median MSE,

$$\text{median}_d \, \frac{1}{N} \sum_{i=1}^{N} (\hat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^2, \quad d = 0, \ldots, 20,$$

when using **T2** and removing any data identified as outlying.

Plots illustrating the same comparison when data sets are contaminated by clusters of outliers are not shown because the median MSE for data sets not trimmed become absurdly large. For sample sizes of $n = 100$, consisting of 10 outliers, the median MSE grows to values in excess of 4 and for data sets with 30 outliers, the median MSE grows to values in excess of 16 for $d \to 20$. The median MSE for non-robust estimates for location have, typically, no upper bound.

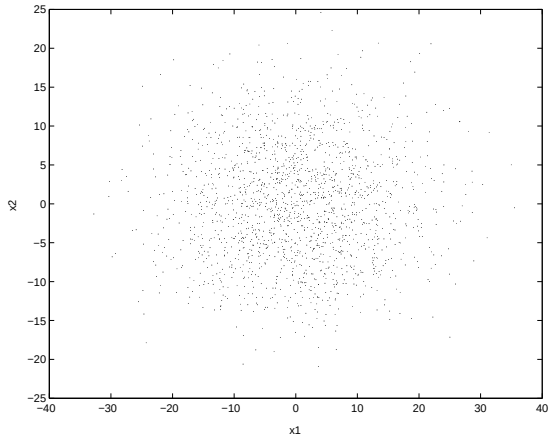


Figure 2.16: $\rho_{12} = \rho_{21} \approx 0$.

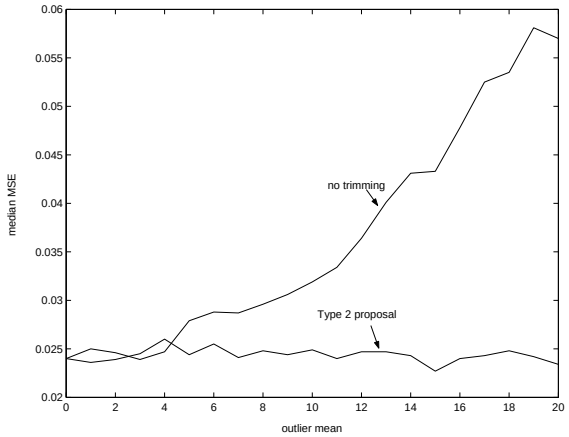


Figure 2.17: One outlier no trimming, $n = 100$, $p = 3$.

2.5 Comparison with Fixed Threshold Methodology

For a formal comparison between the adaptive threshold methodology and other fixed threshold methodologies, which we describe below, we generated a further series of Monte Carlo simulations. The simulations involved corrupting a proportion, $\epsilon = 1/n, 0.1, 0.3$ respectively, of each generated sample's pth-variable. This proportion of the pth-variable was displaced about an increasing range of *outlier means*,

$$d = 1, \ldots, 20.$$

That is, the contaminated proportion of the pth-variable was distributed $N(d, 1)$ with respect to the main sample, distributed $N(\mathbf{0}, \boldsymbol{I}_p)$.

With respect to *each* of the outlier mean displacements d, the median standard square errors

$$\text{median}_d \left\{ (\mathbf{x}_{id} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{id} - \boldsymbol{\mu}) \right\}$$

of the corresponding simulation estimates for location were computed. Noticing that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ since, with the outliers removed, the data sets are ideally generated $N(\mathbf{0}, \boldsymbol{I}_p)$.

This comparison between the new proposal and fixed threshold algorithms has been depicted in a series of Figures 2.18-2.30. In these plots we have 3 algorithms for detecting outliers using a fixed threshold, pre-specified cut-off region, each beginning with an MCD estimate for multivariate location and scale. The plots show the average amount of trimming advised by the algorithm with respect to the average outlier displacement from the majority sample $S_{(1-\epsilon)n} \sim N(\mathbf{0}, \boldsymbol{I}_p)$. For an average outlier displacement of $d < 4$ it may not be surprising to find even the "ideal" algorithm not identifying outliers since this level of displacement should be well within the confidence regions for estimates of a population mean. As the outlier mean increases we should expect to see these algorithms indicating an average trimming amount converging to the true quantity of outliers planted, denoted by the dashed line.

Three Fixed Threshold methods, **FT1, FT2** and **FT3**, were constructed according to the general case (see Rousseeuw and van Zomeren 1990, Hadi 1992, 1994) without correction factors (see Hadi 1994).

For **FT1** the algorithm involved 2 steps:

Step 1: Calculate an MCD estimate for location and scale.

Step 2: Identify , as outliers, any observation with a Mahalanobis distance from this MCD estimate for centroid beyond the cut-off value $\sqrt{\chi^2_{0.975,p}}$.

For **FT2** we have:

Step 1: Calculate an MCD estimate for location and scale.

Step 2: Use the Forward Search from the new proposal until *every* member of the complement to the retained subset is located beyond $\sqrt{\chi^2_{0.975,p}}$, this complement subset is considered outlying.

The last method **FT3** involved:

Step 1: Calculate an MCD estimate for location and scale.

Step 2: Use the Forward Search from the new proposal until *every* member of the complement to the retained subset is located beyond $\sqrt{\chi^2_{1-0.025/n,p}}$, this complement subset is considered outlying.

The cut-off value used in Step 2 of **FT3** makes use of the Bonferroni inequality,

$$P(\bigcap E_i) \geq 1 - \sum P(E_i^c),$$

where E is any event and E^c is the complement of E. This inequality is derived, using DeMorgan's laws (Berry and Lindgren 1996) from the elementary probability result

$$P(E_1 \cup \ldots \cup E_n) \leq \sum_{i=1}^{n} P(E_i).$$

Figures 2.18-2.20 illustrate the success of these 3 fixed threshold algorithms in comparison with the **T2** algorithm for $p = 3$ dimensional data sets of size $n = 100$ contaminated as outlined above. The average trimming imposed by each method should, ideally, be equivalent to the number of outliers planted, again denoted by the dashed line.

Clearly **FT1** and **FT2** are inadequate. **FT2** identifies far too many normally distributed variables as outlying, **FT1** is only slightly better for small proportions of outliers. The Figures 2.18-2.20 show **FT3** is as accurate as **T2** at identifying outliers.

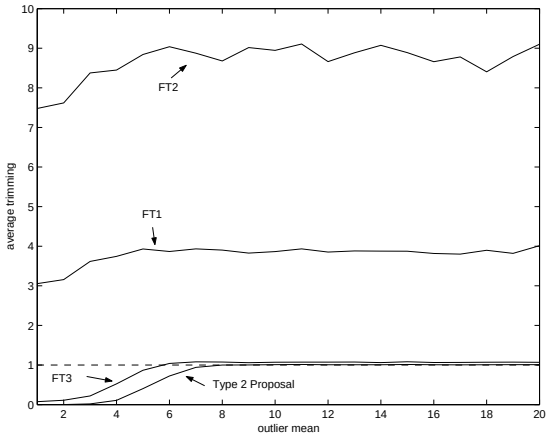


Figure 2.18: **T2** vs Fixed Threshold $n = 100,\ p = 3,\ \epsilon = 0.01$.

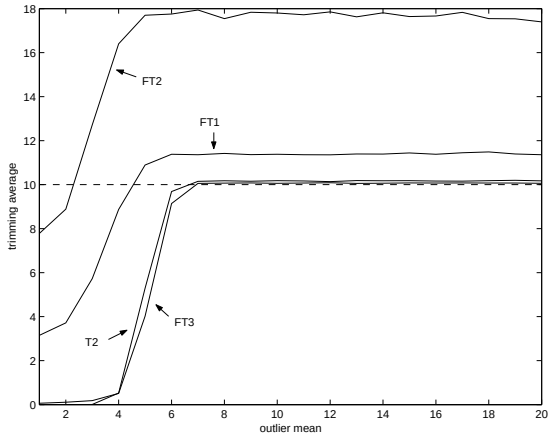


Figure 2.19: **T2** vs Fixed Threshold $n = 100,\ p = 3,\ \epsilon = 0.1$.

Figures 2.21-2.23 depict the results obtained when applying the algorithms to $p = 10$ dimensional data sets of size $n = 500$. The performance of all 4 algorithms reflect the same degree of success as when applied to samples of size $n = 100$, dimension $p = 3$.

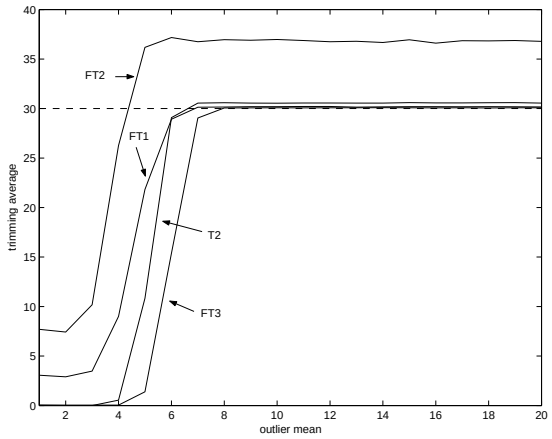


Figure 2.20: **T2** vs Fixed Threshold $n = 100,\ p = 3,\ \epsilon = 0.3$.

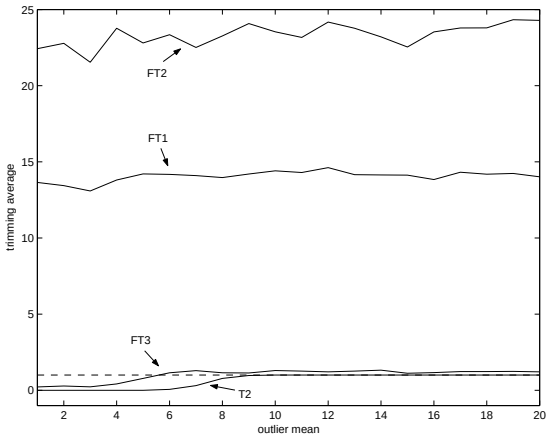


Figure 2.21: **T2** vs Fixed Threshold $n = 500,\ p = 10,\ \epsilon = 0.002$.

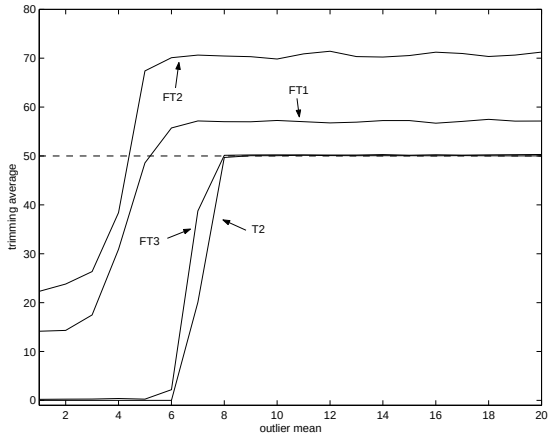


Figure 2.22: **T2** vs Fixed Threshold $n = 500,\ p = 10,\ \epsilon = 0.1$.

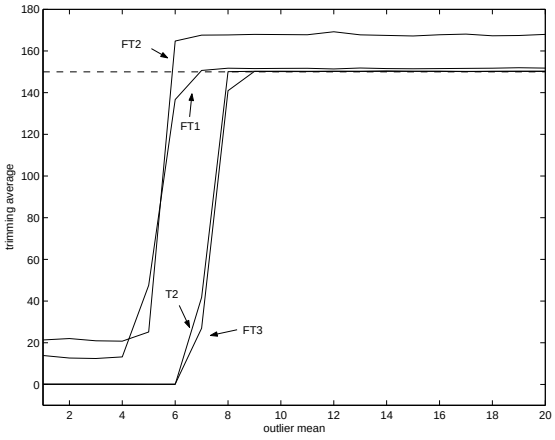


Figure 2.23: **T2** vs Fixed Threshold $n = 500,\ p = 10,\ \epsilon = 0.3$.

Figure 2.24 offers a clearer picture of the advantage of using **T2**, it depicts the trimming average when **FT1**, **FT2**, **FT3** and **T2** are applied to *clean* $p = 3$ dimensional data sets of sizes $n = 100$ through to $n = 1000$. Clean observations are increasingly vulnerable to being identified as outliers by **FT1** and **FT2** as the sample size increases. Figure 2.25 shows **FT3** also followed a similar trend to **FT1** and **FT2** but on a much smaller scale, for samples of size $n = 1000$, dimension $p = 3$, **FT3** is still wrongly identifying observations

as outlying at a level small enough to be acceptable. **T2** *rarely* identifies any clean data as outlying as the sample size increases, the bigger the sample size, the less likely **T2** will identify clean observations as outliers.

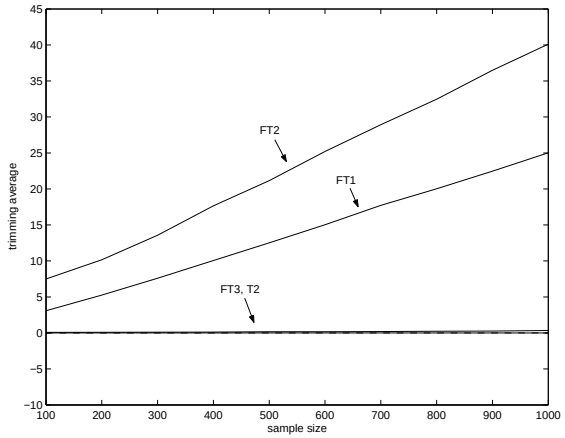


Figure 2.24: **T2** vs Fixed Threshold $n = 100, 200, \ldots, 1000,\ p = 3,\ \epsilon = 0.$

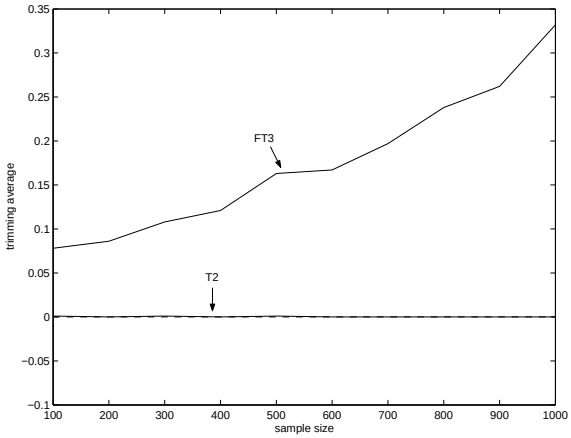


Figure 2.25: **T2** vs **FT3** $n = 100, 200, \ldots, 1000,\ p = 3,\ \epsilon = 0.$

Figures 2.26-2.27 plot the trends when identical comparisons were carried out for clean data sets of size $n = 100$ for an increasing dimension, $p = 2$ through to $p = 10$. The performances of the methodologies mirror their respective performances when applied to the clean data sets increasing in size above.

Due to the impressive performance of **FT3** it was decided to subject this algorithm to more tests in comparison with **T2**. Figure 2.28 shows their performance in detecting *two* clusters of outliers in $p = 3$ dimensional data sets of size $n = 100$. One cluster composed a proportion $\epsilon_d = 0.2$ of the sample with its pth-variable displaced about a shifted mean of $d = 10, \ldots, 30$, a second cluster with an additional proportion, $\epsilon_{d/2} = 0.2$, of the data sets pth-variable displaced $d/2$ with respect to the first. Figure 2.28 show **T2** detects the two clusters much earlier than **FT3**, equivalently when the shifted mean of the clusters is smaller.

Figure 2.29 shows that, when applied to $p = 10$ dimensional data sets of size $n = 500$, **T2** and **FT3** are equally as efficient at detecting two outlying clusters whence a proportion

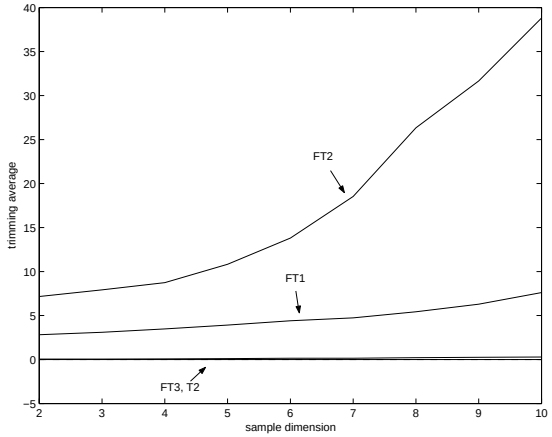


Figure 2.26: **T2** vs Fixed Threshold $n = 100,\ p = 2, 3, \ldots, 10,\ \epsilon = 0.$

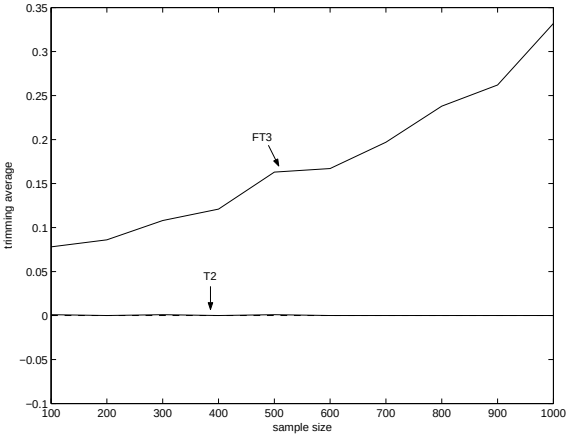


Figure 2.27: **T2** vs **FT3** $n = 100,\ p = 2, 3, \ldots, 10,\ \epsilon = 0.$

$\epsilon_{p\text{th}} = 0.2$ of the pth-variable is distributed $N(d, 1)$ and a proportion, $\epsilon_{(p-1)\text{th}} = 0.2$, of the $(p-1)$th-variable is distributed $N(d/2, 1)$.

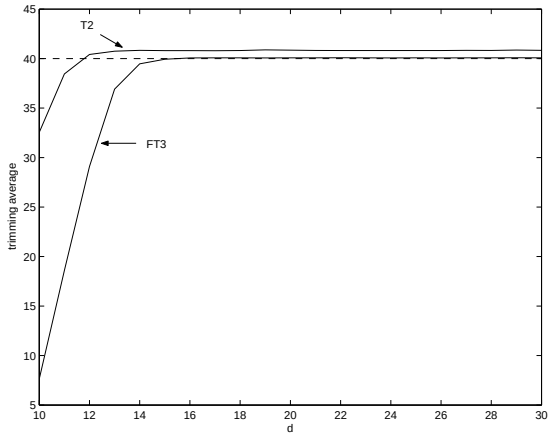


Figure 2.28: **T2** vs **FT3** $n = 100,\ p = 3,\ \epsilon_d = 0.2,\ \epsilon_{d/2} = 0.2.$

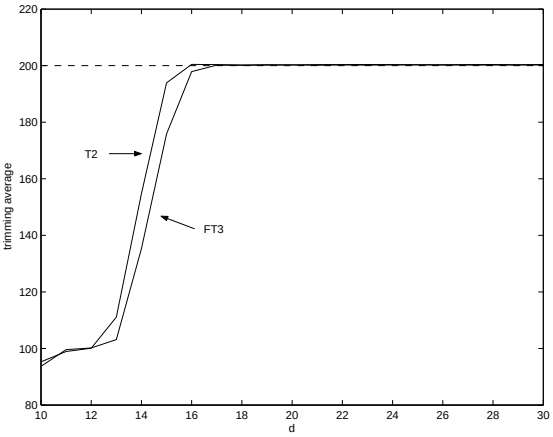


Figure 2.29: **T2** vs **FT3** $n = 500,\ p = 10,\ \epsilon_{p\text{th}} = 0.2,\ \epsilon_{(p-1)\text{th}} = 0.2.$

Figure 2.30 shows a combined plot of the comparison between **T2** and **FT3** at identifying outliers when applied to $p = 10$ dimensional data sets of size $n = 50$. The dashed lines indicate the three different proportions of planted outliers, $\epsilon = 1/n, 0.1, 0.3$, and the dotted line the average trimming of outliers advised by **FT3**. In comparison with the solid line,

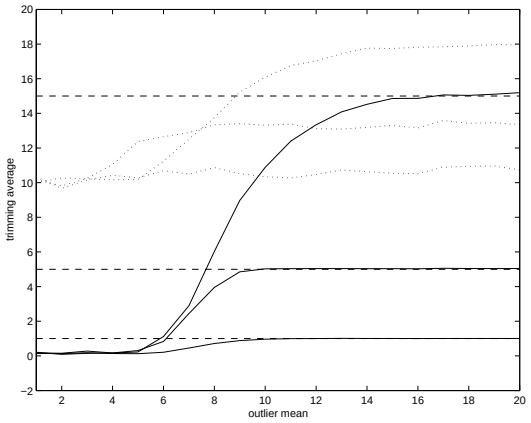


Figure 2.30: **T2** vs **FT3** $n = 50,\ p = 10,\ \epsilon = 0.02, 0.1, 0.3.$

describing the performance of **T2**, we can see **FT3** is identifying too many observations as outliers.

2.6 The T2 Algorithm - further deliberations

Returning to the phenomena of multiple minima occurring for an $\alpha > 0$, it was initially thought to choose that α corresponding to the *minimum of any minima*, $\min_i(\mathsf{m}_i)$, for $\alpha > 0$. This proved fine and always coincided with choosing that minimum corresponding with the largest $\alpha > 0$, m_j. This relationship between the subset, $S_{\gamma_{\min_i(\mathsf{m}_i)}}$, corresponding to the *minimum minima* for $\alpha > 0$ being equivalent to the subset, $S_{\gamma_{\mathsf{m}_j}}$, corresponding to the greatest $\alpha > 0$ for which a minimum occurred was in fact violated, thus

$$S_{\gamma_{\min_i(\mathsf{m}_i)}} \neq S_{\gamma_{\mathsf{m}_j}},$$

by the Monte Carlo series of samples contaminated by 2 clusters, constructed as for the results depicted in Figures 2.28-2.29. For those simulations the subset $S_{\gamma_{\mathsf{m}_j}}$ was chosen for the **T2** subset of retained data. Figures 2.31-2.32 highlight the important difference between choosing that subset of data which corresponds to the minimum of any minima occurring for an $\alpha > 0$, when using the new proposal, versus choosing that minima for any

$\alpha > 0$ corresponding to the smallest subset of retained data. In most outlying cases we will only find *one* minimum to the objective function. When multiple minima do occur the minimum of the multiple minima occurring for $\alpha > 0$ will usually correspond to the minimum subset of retained data, that is the minima which trims the *most* data. Figures 2.31-2.32 are drawn up for the response of **T2** when applied to Monte Carlo samples equivalent to those used for the simulation results depicted Figures 2.28-2.29 and show there is an important difference between the two choices. There is very strong evidence here to suggest that, in the event of multiple minima for $\alpha > 0$, choosing that minimum corresponding to the smallest subset of retained data is the perhaps the most efficient target. In fact Figure 2.32 shows, when seeking the minimum of the multiple minima occurring for an $\alpha > 0$, **T2** only ever identifies *one* cluster.

This outcome could perhaps have been predicted when we return to the theoretical argument for seeking minima to the objective function. The idea that unless a data set is uni-modal there will occur minima when uni-modal subsets are assessed using (2.6) for $n < 30$ and (2.7) otherwise. The phenomenon of multiple minima will be fully examined in Chapter 4.

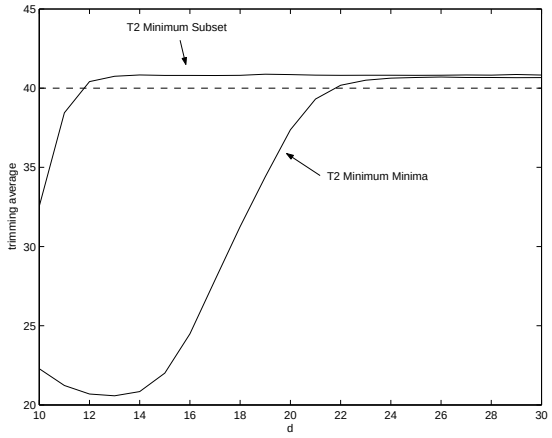


Figure 2.31: $S_{\min_i(\mathsf{m}_i)} \neq S_{\mathsf{m}_j}$ $n = 100$, $p = 3$, $\epsilon_d = 0.2$, $\epsilon_{d/2} = 0.2$.

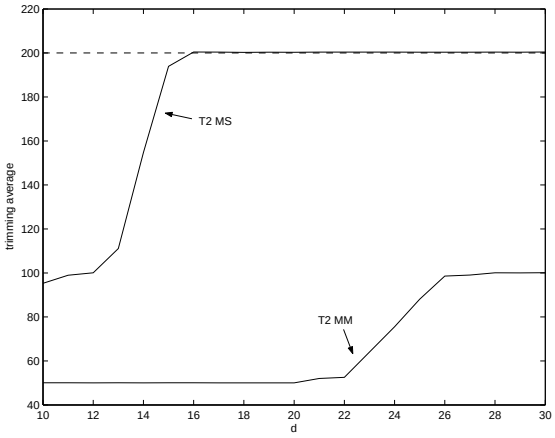


Figure 2.32: $S_{\min_i(\mathsf{m}_i)} \neq S_{\mathsf{m}_j}$ $n = 500$, $p = 10$, $\epsilon_{p\text{th}} = 0.2$, $\epsilon_{(p-1)\text{th}} = 0.2$.

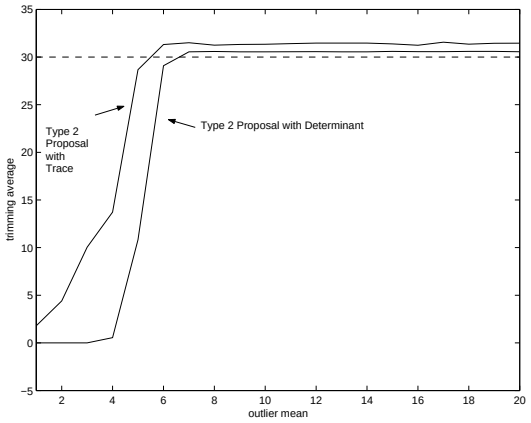


Figure 2.33: Determinant vs Trace $n = 100,\ p = 3,\ d = 0, \ldots, 20$.

2.6.1 Determinant vs Trace

The **T1** and **T2** proposal use the value of a determinant of the measure for the asymptotic variance of an estimate for location. While the determinant of the asymptotic covariance matrix of the MCD estimate for location is one measure that can be used, there exist many others, one of which is the *trace*. Both measures reduce to the usual variance for $p = 1$ (Johnson & Wichern 1998).

Figure 2.33 depicts the comparison between applying **T2** using the determinant and applying **T2** using the trace of the covariance measure on $p = 3$ dimensional data sets of size $n = 100$. It is clear, from Figure 2.33, that when using the trace the objective function has a tendency to overtrim for, in this case $\epsilon = 0.3$, one cluster shifted from the main sample mean over a range of outlier means $d = 0, \ldots, 20$.

2.7 Gervini comparison

An alternative algorithm involving the adaptive approach, whence the threshold value is determined uniquely for each data set by the algorithm, is the Adaptive Reweighted Es-

timator (Gervini 2003). From various starting points, minimum volume ellipsoid (MVE), MCD and S-estimate for multivariate location and scale, Gervini assesses each observation's Mahalanobis distance against threshold values $u \geq \eta$, where $\eta = \chi^2_{1-\alpha,p}$ for an arbitrary, small α. Signifying the value of the χ^2_p distribution at u by $G_p(u)$, this algorithm calculates an adaptive threshold value from

$$\alpha_n = \sup_{u \geq \eta} \{G_p(u) - G_n(u)\}^+$$

where $G_n(u)$ is one minus the proportion of observations in the *sample* data lying beyond u. Of course negative differences are ignored since $G_n(u) > G_p(u)$ infers the expectation of the existence of such a normally distributed point(s). Assume we have a sample of size one-thousand and points $u_1 < u_2 < u_3$ such that $G_p(u_1) = 0.98$, $G_p(u_2) = 0.99$ and $G_p(u_3) = 0.9999$ while $G_n(u_1) = 0.90$, $G_n(u_2) = 0.999$ and $G_n(u_3) = 0.999$. Then our interpretation would be that *a proportion* $\epsilon = (0.9999 - 0.999) = 0.0009$ *of sample data* is extraneous with respect to the point u_3. But at point u_2 no more sample points appear outlying, whilst at u_1 we see that this $\epsilon = 1 - G_n(u_2) = 0.001$ proportion is indeed a member of the set of extraneous points. At u_1 a proportion $\epsilon = 0.10$ of the sample data lies beyond the "population" quantity corresponding to $\chi^2_p = 0.98$. Therefore, in this case, if

$$G_p(u_1) - G_n(u_1) = \sup_{u \geq \eta} \{G_p(u) - G_n(u)\}^+$$

then the trimming proportion would be $\alpha_n = 0.08$.

This algorithm has a tendency to trim *clean* data sets even for sample sizes as high as $n = 500$ (Gervini 2003). When the **T2** proposal is applied to clean data sets of size $n \geq 100$ trimming of data occurs very rarely, see Table 2.1. Hence the estimates for location and scale of clean samples are equal to the sample mean and covariance matrix.

For a direct comparison with Gervini (2003) we documented the errors in the estimates for multivariate location and scale when **T2** was applied to $p = 3$ and $p = 10$ dimensional data sets of sizes $n = 50$ and $n = 500$. The level of pth-variable contamination in the ensuing four sample types was $\epsilon = 0.1$ and $\epsilon = 0.2$ (see Gervini 2003).

To assess the accuracy of the estimates for location, the *maximum* median MSE of $\hat{\boldsymbol{\mu}}_\alpha$

with respect to the size, d, of mean displacement of the outlying proportion was calculated. Meaning that in this case, the medians for *every* outlier mean displacement d was examined and the maximum of these medians was recorded. The errors in the estimate for scale were calculated, again with respect to d, using the maximum median Log Condition Number (LCN) of $\hat{\mathbf{\Sigma}}_\alpha$, the covariance matrix of the retained data. The Condition Number is defined as **the ratio of the largest eigenvalue to the smallest** (Brown 2004). When dealing with a covariance matrix this condition number describes the shape of the data since the value represents the ratio between the variance explained by the most dominant variable and the variance explained by the least dominant. This eigenvalue ratio should therefore be close to 1 for clean randomly generated data.

The maximum, with respect to d, median MSE of $\hat{\boldsymbol{\mu}}_\alpha$ and LCN of $\hat{\mathbf{\Sigma}}_\alpha$ when compared with an assumed location and LCN for an *uncontaminated* data set was recorded in Table 2.11. We discuss the results in comparison with Gervini (2003) in the next paragraph.

The maximum of the median standard errors, regarding the location estimate, occurred for $4 \leq d \leq 7$ and are slightly greater in comparison to Gervini (2003). For any $d \geq 10$ the outliers were nearly always identified, thus the results appeared to converge to the expected parameter estimates of a *clean* data set. The median SE results, using **T2**, for the simulations conducted for $p = 10$ dimensional data sets were slightly greater than those for the Gervini Adaptive Reweighted Estimate starting with the MCD estimate for location and scale. For example the maximum median SE's, for $p = 10$ dimensional samples, arrived at by Gervini (2003) were, on average, 0.51. The results, using the new proposal for the error in scale estimate, for the maximum median LCN's were slightly smaller than those for the Gervini Adaptive Reweighted Estimator when applied to 3 dimensional samples and considerably smaller for the $p = 10$ dimensional data sets. Gervini's corresponding average LCN for $p = 3$ dimensional samples was 1.50 whilst for $p = 10$ dimensional samples, of size $n = 50$, the average LCN was 3.85.

Simulations comprising p-dimensional data sets with a proportion $\epsilon = 0.2$ of their pth-variable distributed $N(0, 50)$ were investigated under a similar criteria. These generated

n	$\boldsymbol{\rho}$	p_t	$\bar{\alpha}$
20	-0.95	0.033	0.0090
	-0.50	0.035	0.0108
	0	0.060	0.0155
	+0.50	0.040	0.0119
	+0.95	0.038	0.0092
50	-0.95	0.003	0.0008
	-0.50	0.008	0.0015
	0	0.009	0.0014
	+0.50	0.005	0.0006
	+0.95	0.007	0.0012
100	-0.95	0.003	< 0.0001
	-0.50	0.003	< 0.0001
	0	0.003	< 0.0001
	+0.50	0.002	< 0.0001
	+0.95	0.001	< 0.0001

Table 2.10: $\rho_{12} = \rho_{21} = \boldsymbol{\rho}$.

Table 2.11: Errors of location and scatter estimates for shifted normal.

Error in location estimate				*Error in scale estimate*			
n	p	ϵ	maximum median SE	n	p	ϵ	maximum median LCN
50	3	0.1	0.1848	50	3	0.1	1.1136
		0.2	0.65			0.2	1.5225
	10	0.1	0.5134		10	0.1	2.408
		0.2	1.7089			0.2	2.9263
500	3	0.1	0.2191	500	3	0.1	1.1773
		0.2	0.9657			0.2	1.6418
	10	0.1	0.4666		10	0.1	1.8951
		0.2	1.9266			0.2	2.3977

Table 2.12: Errors of location and scatter estimates for amplified variance.

Error in location estimate			*Error in scale estimate*		
n	p	median SE	n	p	median LCN
50	3	0.0604	50	3	0.6912
	10	0.2348		10	1.9005
500	3	0.0057	500	3	0.2174
	10	0.0239		10	0.5649

samples had contaminants centred about the same mean as the clean data but possessing an amplified variance and the median standard error of the calculated estimates for location were calculated. The results, see Table 2.12, are excellent for $n = 500$ and regarding $n = 50$, as good as the results when the Gervini Adaptive Reweighted Estimate was applied to data sets with the same proportion, $\epsilon = 0.2$, of the pth-variable shifted.

The **T2** proposal was also applied to Cauchy distributed data sets of size $n = 50$ and $n = 500$, generated in both 3 and 10 dimensions, for further comparison with Gervini (2003). Table 2.13 contains the relative median standard error of the estimate with respect to the Cauchy maximum likelihood estimate, MLE, for location, (see Gervini 2003), and shows that the new proposal has yielded estimates for location very similar to the MLE.

Investigating the shape of the estimate for scale, we chose to calculate the relative median LCN with respect to the Cauchy MLE for the LCN echoing, again, the analysis of Gervini (2003). The new proposal estimate for scale produced relative values of $\approx 70\%$ to the median of the LCN's derived from the Cauchy MLE. These results for Cauchy data are as good as those found in Gervini (2003).

It is interesting to notice that when trimming any data set of planted outliers the median SE of estimates for location (Gervini 2003) will not be impacted by *over*trimming. If clean data is identified as outlying, along with the contaminants, a median SE will not be affected greatly.

Table 2.13: Errors in Cauchy estimation with respect to Cauchy MLE.

Error in location estimate			*Error in scale estimate*		
n	p	Relative median SE	n	p	Relative median LCN
50	3	0.9552	50	3	0.7072
	10	0.9298		10	0.6839
500	3	0.9354	500	3	0.7258
	10	0.9367		10	0.7069

2.8 Online data sets

Figure 2.34 depicts the bivariate plot for the relationship between the size of the acorn and the geographic range for various Atlantic and California oak tree species from Aizen and Patterson (1990), Journal of Biogeography, volume 17, p. 327-332. Of the 39 observations that denoted by the triangle at (13,7.1) Quercus tomentella Engelm (California) is said to be the outlier here but the **T2** proposal did not trim this observation. The sole observation trimmed by the new proposal is denoted by the cross at (690,17.1) where a *global* minimum was found and indeed it does appear that this trimmed observation is the most outlying.

It has to be noted, when inspecting Figure 2.35, that the number of subsets minimizing (2.7) lead to the conclusion that the data set is *erratically* distributed, a messy sample.

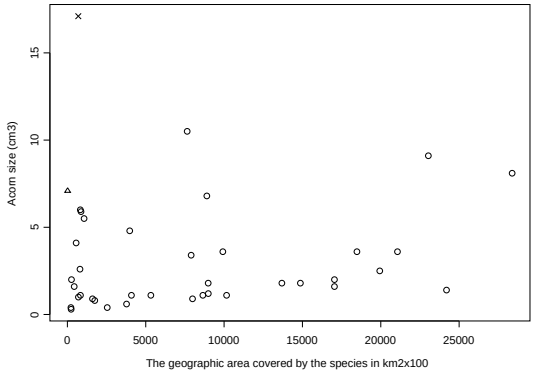


Figure 2.34: Acorn data set.

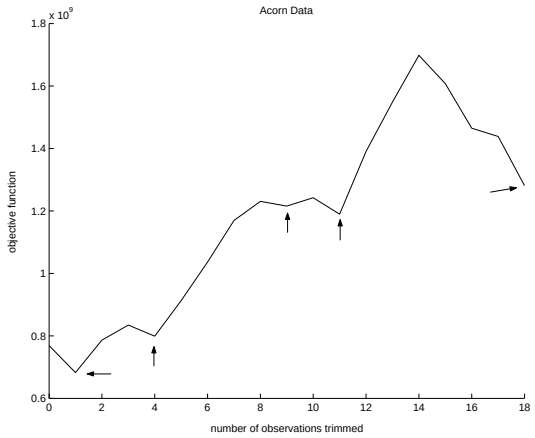


Figure 2.35: Minima occurring.

Figure 2.36 plots the relationship between the age of Chief Executive Officer and his or her

corresponding salary for $n = 60$ small firms in 1993 according to Forbes magazine (Forbes, November 8, 1993, "America's Best Small Companies"). These were firms with annual sales of more than \$5 million and less than \$350 million. No observation was deemed an outlier by Forbes but one may regard the observation marked with a cross (57,1103) suspiciously outlying.

Figure 2.37 concerns bivariate data for the distances two identical footballs were kicked by a *novice punter* on a windless day at The Ohio State University's athletic complex. The only difference between the balls was one was filled with air the other helium. According to Lafferty, M. B. (1993), "OSU scientists get a kick out of sports controversy", the Columbus Dispatch (November, 21, 1993), B7, there were two *types* of possible outliers amongst the $n = 39$ observations, those kicks less than 15 yards and kicks less than 20 yards. **T2** was applied to both these data sets and did not identify any observation as outlying.

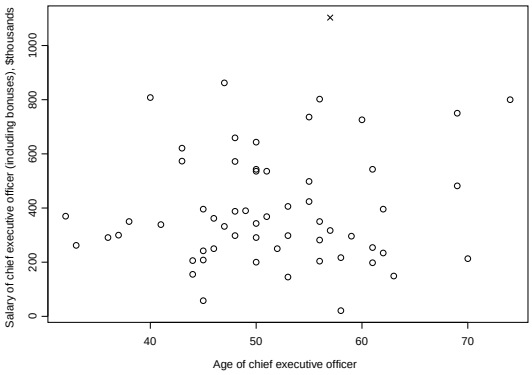


Figure 2.36: CEO data set.

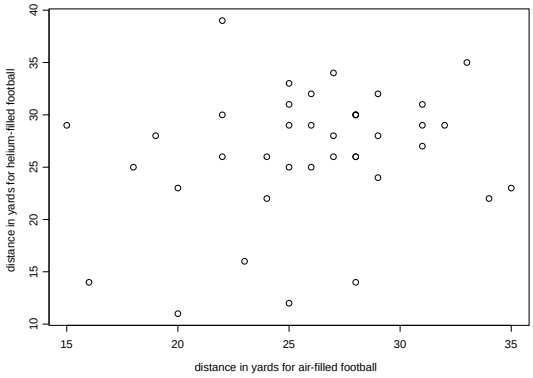


Figure 2.37: Football's kicked data set.

Figure 2.38 care of J.M. Hunter, "Need and Demand for Mental Health Care: Massachusetts 1854." The Geographic Review, 77:2 (April 1987), pp 139-156. The data are from an 1854 study involving the percentage of lunatics cared for at home and distance to the nearest health centre. The observation 13 at (77,25), Nantucket, is known to be an outlier and the new proposal **T1**, implemented because of the small sample size, confirmed this observation and observation 8 at (4,6), Suffolk, as outliers. An inspection of the

plot is an assurance that Suffolk indeed appears suspiciously outlying. The **T2** proposal identified these two points as outliers as well as observation 1 (97,77), which is arguably extreme.

Figure 2.39 plots the size of the **T1** objective function in relation to the number of removed observations when identified as outlying. Notice how borderline the difference was between the value of **T1** for $\alpha = 2/n$ and $\alpha = 3/n$, the values were $V(2/14, F_n) = 3.9411 \times 10^5$ and $V(3/14, F_n) = 4.0170 \times 10^5$ respectively.

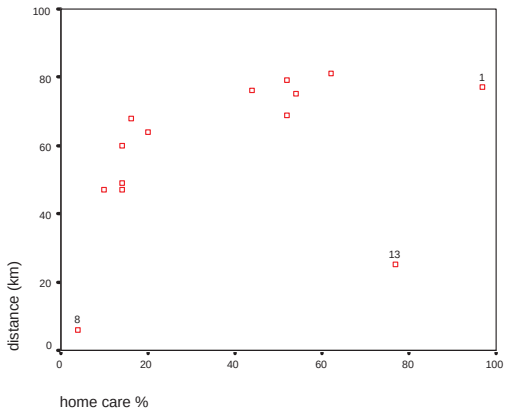


Figure 2.38: Massachusetts lunatics 1854.

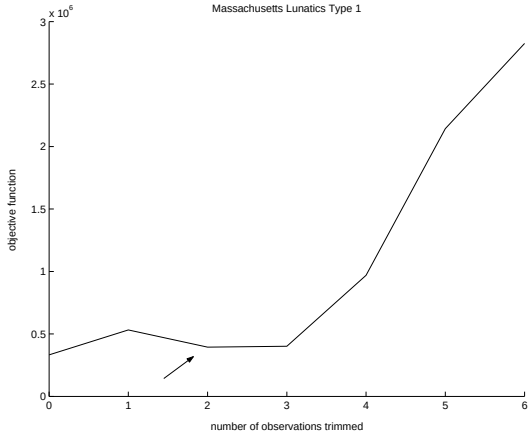


Figure 2.39: Minimum occurring.

Figure 2.40 plots the leading quarterback salary versus the total team salary for $n = 28$ football teams in the American Football Conference (AFC) and National Football Conference (NFC) of the National Football League(NFL) in the 1991 season as reported by the Associated Press. The observations at (30131,3500), Steelers, and (23074,300), Bears, are considered potential outliers. **T1** did not judge these two observations as outliers.

Figure 2.41 concerns Branch Rickey's set of $n = 25$ outstanding hitters in baseball over the period 1920 to 1950 from "Good-bye to Some Old Baseball Ideas", Life Magazine, August 2, 1954. The question remains is Babe Ruth (481,271) an outlier? The new proposal, **T1**, did not judge the Babe Ruth figures an outlier for this data set.

Figure 2.42 is a scatter plot from a 1965 report A.J. Lea (1965), "New Observations on Distribution of Neoplasms of Female Breast in Certain Countries", (British Medical Journal, 1, 488-490), examined the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women. The subjects were residents of $n = 16$ different regions of Great Britain, Norway, and Sweden. According to a simple regression of mortality index on temperature the cross at (31.8,67.3) is an outlier but **T1** did not agree with this inference.

Worthy of note is the fact that when the more sensitive **T2** was applied to the data sets described in Figures 2.40, 2.41 and 2.42 no outliers were identified.

Figure 2.43 depicts percent changes in manpower and seasonally adjusted changes in weekly auto thefts for the $n = 23$ precincts in New York City from a base period of 27 weeks in 1966 to an experimental period of 58 weeks in late 1966 and 1967 published by S.J. Press, "Some Effects of an Increase in Police Manpower in the 20th Precinct of New York City"", The New York City Rand Institute, R-704-NYC, October 1971. Precinct number 20 at (39.4,-2.65) for which manpower assigned was increased by about 40 percent is an obvious outlier here and was identified as such by **T1** with a global minimum of (2.6) occurring when this observation was trimmed.

Figures 2.40-2.43 are examples of small, bivariate data sets and how extreme an observation needs to be to warrant outlier status using the new proposal.

The next two plots, Figures 2.44-2.45, present brilliant examples of solitary outlier observations that will remain undetected if one uses a regression analysis since they act as good leverage points.

Figure 2.44 is the bivariate scatter plot the care of U.S. Department of Commerce, Bureau of the Census, Government Finances in 1960, Census of Population, 1960, Census of Manufactures, 1958, Statistical and Abstract of the United States, 1961. U.S. Department of Agriculture, Agricultural Statistics, 1961. U.S. Department of the Interior, Minerals Yearbook, 1960. The plot specifically exhibits the relationship between the Economic

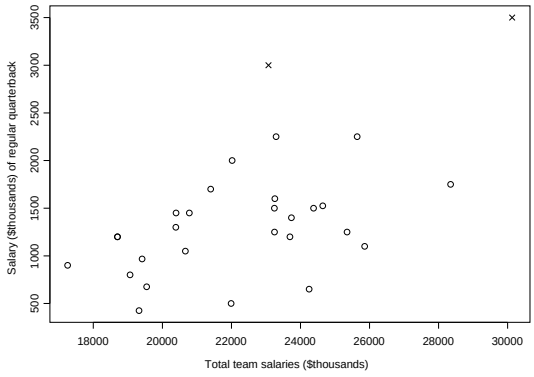


Figure 2.40: Quarterback data set.

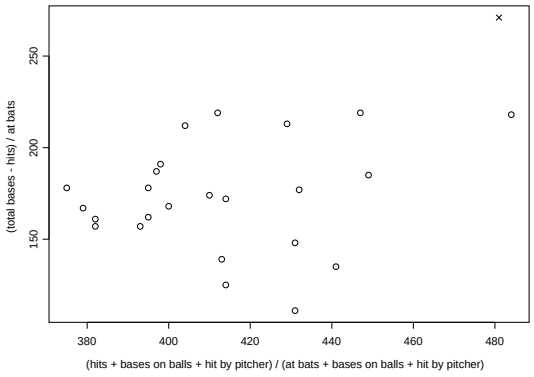


Figure 2.41: Babe Ruth data set.

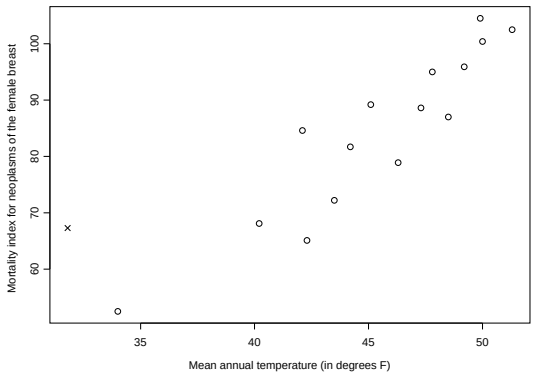


Figure 2.42: Breast Cancer data set.

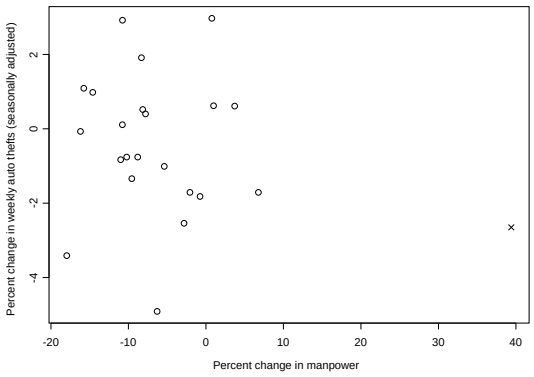


Figure 2.43: New York Police data set.

Ability Index and Expenditures. The observation corresponding to Nevada at (205,421), marked with a cross, is the known outlier here and this was confirmed by the **T2** proposal forcing a global minimum of (2.7) when removed.

Figure 2.46 plots the average public teacher pay versus the spending on public schools per pupil, in 1985, for the 50 states of America and the District of Columbia as reported by the Albuquerque Tribune. The outlier denoted by the cross corresponded with Alaska and when trimmed, forced a global minimum of **T2** which, again, highlights the power of the

new proposal, as in the previous example, these outlying observations would not have been detected if regression analysis was used to explain the relationships. This is an expected consequence, when using an algorithm assessing measures of covariance determinants for location estimates, as opposed to tracking the sizes of residuals using a regression fit. A good leverage point, no matter how extreme, will not possess a large residual with respect to the majority data. The measure calculated, when using **T1** and **T2**, is impacted by displacement from a centroid estimate with respect to a scale estimate and so leverage is not an issue.

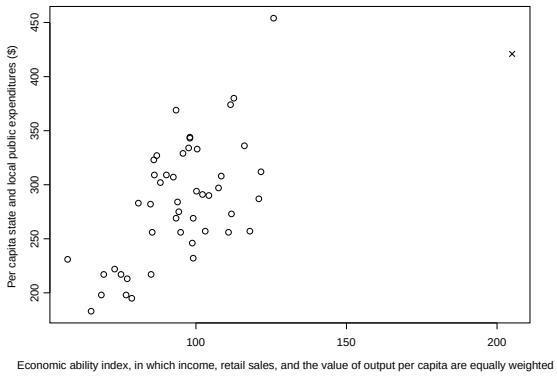


Figure 2.44: State Spending data set.

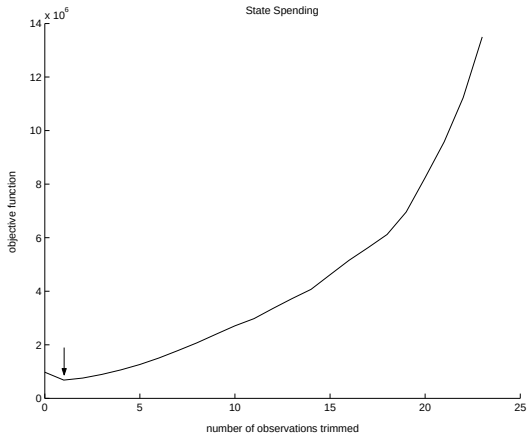


Figure 2.45: Minimum occurring.

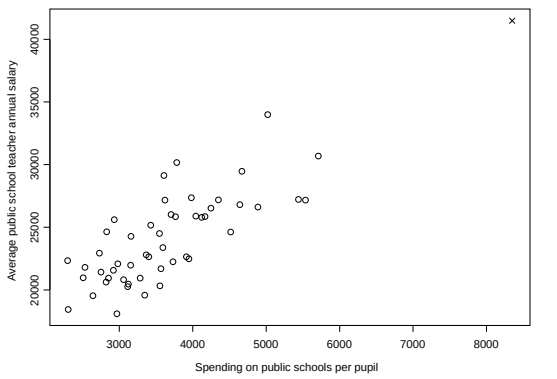


Figure 2.46: Teachers Pay data set.

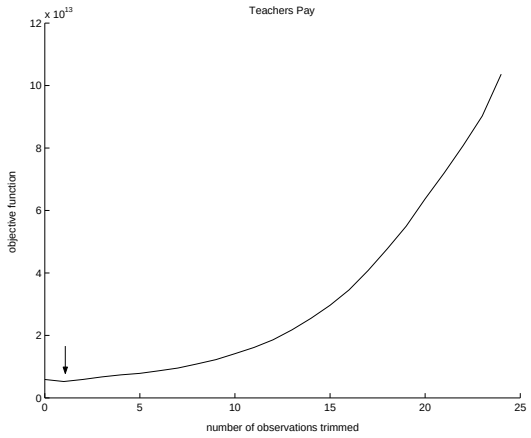


Figure 2.47: Minimum occurring.

The data used for the next two plots, Figures 2.48-2.9, appeared in the Wall Street Journal in March 1, 1984 . Advertisements were selected by an annual survey conducted by Video Board Tests, Inc., a New York ad-testing company, based on interviews with 20,000 adults who were asked to name the most outstanding TV commercial they had seen, noticed, and liked. The retained impressions were based on a survey of 4,000 adults, in which regular product users were asked to cite a commercial they had seen for that product category in the past week. Those impressions retained were tabulated per million and compared with the TV advertising budget corresponding to the firm screening the ad. This data set produced some exciting results, even though the sample size is only $n = 21$, since *two* local minima occurred away from $\alpha = 0$. The observations represented by the filled in squares were excluded from the data set whence the first local minimum of (2.6) occurred and the observations represented by triangles were added to these three forcing a second local minima. Now if we take the $\alpha > 0$ corresponding to that minimum associated with the smallest subset, $S_{\gamma_{\mathsf{m}_j}}$, of retained data as our correct trimming proportion we trim all 8 observations denoted by the squares and triangles. One may safely dispute the normality of such a data set but Figure 2.49 illustrates that when the 3 obvious outliers signified by the filled in squares were removed, this coincided with the minimum of the two minima, $\min_i(\mathsf{m}_i)$, occurring for $\alpha > 0$. This TV adds data set may well be considered as composing *two* clusters, denoted by the crosses and the triangles, and *three* stray points denoted by the squares.

The problem posed in the last example on TV Adds is again an issue in the next $p = 3$ dimensional example of a real data set with known outliers obtained from the Data and Storage Library at http://lib.stat.cmu.edu/DASL/DataArchive.html. Figures 2.50-2.55 picture the 3-dimensional plots concerning a national sample of 6000 households with the main worker earning less than \$15,000 annually in 1966 (D.H. Greenberg and M. Kosters, "Income Guarantees and the Working Poor", The Rand Corporation (R-579-OEO),

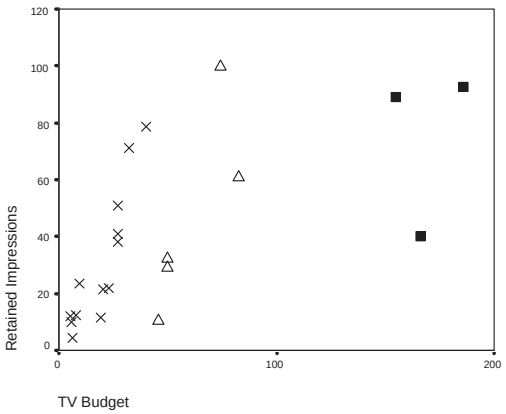


Figure 2.48: TV adds data set.

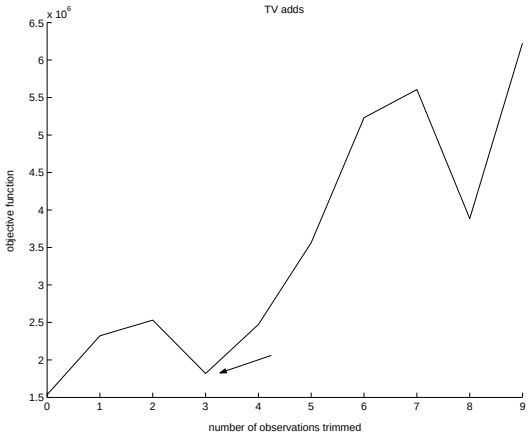


Figure 2.49: Multiple minima occurring.

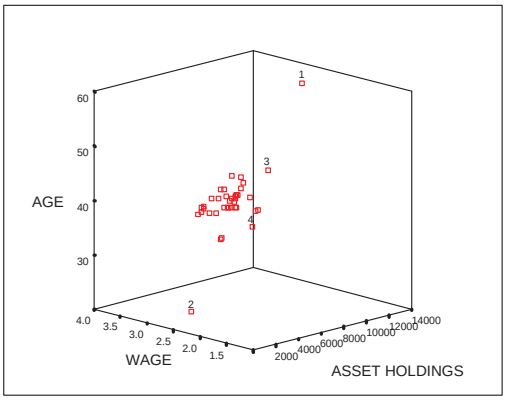


Figure 2.50: Wages hours perspective 1.

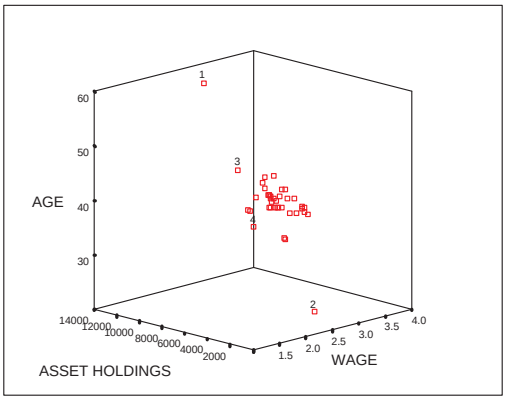


Figure 2.51: Wages hours perspective 2.

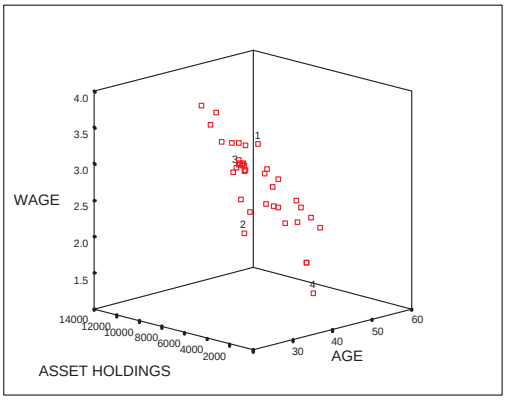


Figure 2.52: Wages hours perspective 3.

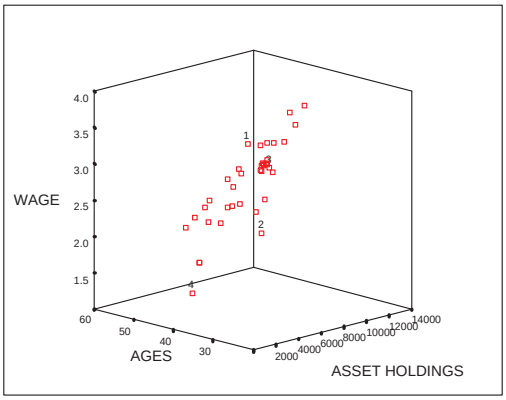


Figure 2.53: Wages hours perspective 4.

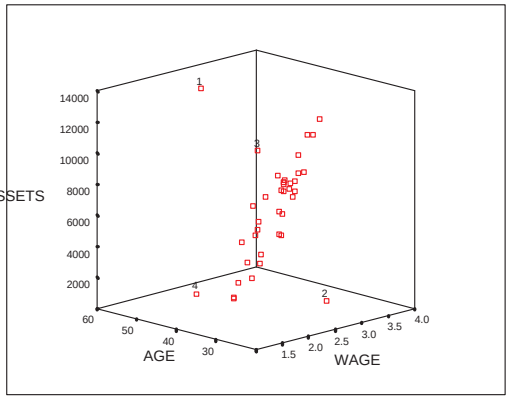


Figure 2.54: Wages hours perspective 5.

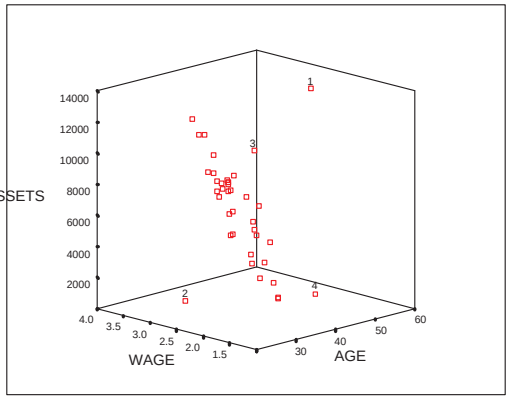


Figure 2.55: Wages hours perspective 6.

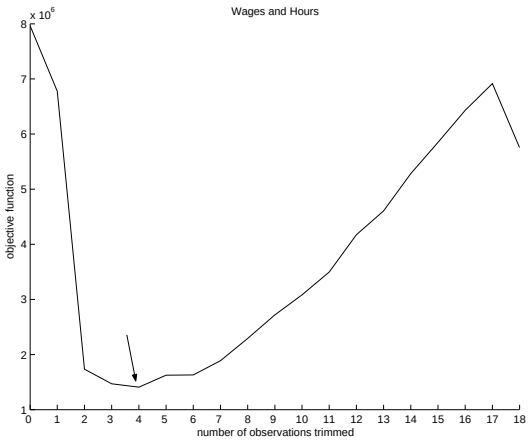


Figure 2.56: Wages Hours Minima.

December, 1970). These 6000 households were divided into 39 demographic subgroups for an analysis of the relationship between average asset holdings, average age and average hourly wage. A three-variable set of predictors found from stepwise regression includes average asset holdings, age and average hourly wage. Average hours, average wages and average asset holdings are positively inter-related. The residuals from the three-predictor regression show one influential outlier at wage=1.42,asset=1866 and age=40.6, marked by the number 4. When **T2** was applied to this data set a *global* minimum occurred when this observation and the three observations marked 1,2 and 3 at (2.79,12710,57.7), (2.51,1632,22.4) and (2.79,9658,43.4) respectively were trimmed.

Figure 2.56 displays the global minimum occurring at $\alpha = 4/39$ when **T2** was applied to all the subsets selected by the Forward search algorithm. It can also be seen that there also occurred a local minimum when 18 of the 39 observations were removed. Should we take this minimum, for the much greater α, as the correct trimming proportion?

The TV Adds and Wages Hours data sets, the simulations for clustered data sets and the consequent presence of multiple minima really indicate a fundamental *non*-normality associated with the data set. These data sets are not *uni*-modal. When multiple minima occur, an event which is increasingly rare as the sample size increases, one might suspect the data set may be clustered. When the minimum of the multiple minima does not

coincide with the minimum corresponding with the greatest α, so that

$$S_{\min_i(\mathsf{m}_i)} \neq S_{\mathsf{m}_j}$$

holds, we can expect the data set is not uni-modal. It has been established, at least empirically, that when **T2** is applied to uni-modal data sets, of size $n \geq 100$, observations are rarely identified as outlying. So we may need to clean the data set of those outliers corresponding to $S_{\min_i(\mathsf{m}_i)}$, then re-apply the new proposal, which in most cases will result in $S_{\min_i(\mathsf{m}_i)} = S_{\mathsf{m}_j}$ for what is now a *trimmed* sample. The few instances where this will not be realized is when the minima corresponding to the original S_{m_i} disappears. As mentioned earlier, this phenomena will be fully examined in Chapter 4.

2.8.1 Cricket Batting Data

The last of the examples using real data sets concern the Career Batting Figures for the top 90 Australian and England cricketers, up to December 31st, 2003. The data set is 4 dimensional and has been chosen because of the extraordinary batting figures of Sir Donald Bradman. The 4 variables chosen for this application of the **T2** proposal were the number of innings played, number of fifties and hundreds scored and the number of runs amassed by these top 90 batsmen.

When applying **T2** there occurred a global minimum at $\alpha = 1/90$ which corresponded to the subset of data with *only* Bradman's figures expelled. This satisfies our methods criteria to consider Bradman's figures a solitary outlier for this 4 dimensional data set.

Plots of the size of the objective function (2.7) and the corresponding number of outliers detected is represented by Figures 2.57-2.58, this minimum occurring when Bradman's figures were removed.

For two 3 dimensional examples using these batting figures we choose innings played, fifties scored and runs amassed in the first example and can see Bradman marked with

Name	x_1	x_2	x_3	x_4	Name	x_1	x_2	x_3	x_4
AR Border	265	27	63	11174	DL Amiss	88	11	11	3612
SR Waugh	258	32	49	10807	AW Greig	93	8	20	3599
GA Gooch	215	20	46	8900	AR Morris	79	12	12	3533
AJ Stewart	235	15	45	8463	EH Hendren	83	7	21	3525
DI Gower	204	18	39	8231	C Hill	89	7	19	3412
G Boycott	193	22	42	8114	GA Hick	114	6	18	3383
ME Waugh	209	20	47	8029	GM Wood	112	9	13	3374
MA Atherton	212	16	46	7728	FE Woolley	98	5	23	3283
MC Cowdrey	188	22	38	7624	KWR Fletcher	96	7	19	3272
MA Taylor	186	19	40	7525	ME Trescothick	81	5	21	3175
DC Boon	190	21	32	7422	VT Trumper	89	8	13	3163
WR Hammond	140	22	24	7249	AC Gilchrist	68	9	16	3159
GS Chappell	151	24	31	7110	MP Vaughan	71	10	8	3118
DG Bradman	**80**	**29**	**13**	**6996**	CC McDonald	83	5	17	3107
L Hutton	138	19	33	6971	AL Hassett	69	10	11	3073
KF Barrington	131	20	35	6806	KR Miller	87	7	13	2958
RN Harvey	137	21	24	6149	WW Armstrong	84	6	8	2863
DCS Compton	131	17	28	5807	GR Marsh	93	4	15	2854
RT Ponting	117	20	21	5749	KR Stackpole	80	7	14	2807
GP Thorpe	151	12	33	5552	NC O'Neill	69	6	15	2779
N Hussain	162	13	30	5430	M Leyland	65	9	10	2764
JB Hobbs	102	15	28	5410	GN Yallop	70	8	9	2756
KD Walters	125	15	33	5357	SJ McCabe	62	6	13	2748
IM Chappell	136	14	26	5345	C Washbrook	66	6	12	2569
MJ Slater	131	14	21	5312	GS Blewett	79	4	15	2552
WM Lawry	123	13	27	5234	BL D'Oliveira	70	5	15	2484
IT Botham	161	14	22	5200	DW Randall	79	7	12	2470
JH Edrich	127	12	24	5138	W Bardsley	66	6	14	2469
TW Graveney	123	11	20	4882	WJ Edrich	63	6	13	2440
JL Langer	116	16	20	4873	TG Evans	133	2	8	2439
RB Simpson	111	10	27	4869	LEG Ames	72	8	7	2434
IR Redpath	120	8	31	4737	MR Ramprakash	92	2	12	2350
AJ Lamb	139	14	18	4656	W Rhodes	98	2	11	2325
H Sutcliffe	84	16	23	4555	WM Woodfull	54	7	13	2300
PBH May	106	13	22	4537	DR Martyn	59	5	15	2292
ER Dexter	102	9	27	4502	TE Bailey	91	1	10	2290
KJ Hughes	124	9	22	4415	PJP Burge	68	4	12	2290
MW Gatting	138	10	21	4409	SE Gregory	100	4	8	2282
ML Hayden	83	17	13	4391	MJK Smith	78	3	11	2278
APE Knott	149	5	30	4389	SK Warne	146	0	8	2238
IA Healy	182	4	22	4356	R Benaud	97	3	9	2201
RA Smith	112	9	28	4236	CG Macartney	55	7	9	2131
MA Butcher	114	8	17	3790	WH Ponsford	48	7	6	2122
RW Marsh	150	3	16	3633	PE Richardson	56	5	9	2061
DM Jones	89	11	14	3631	RM Cowper	46	5	10	2061

Table 2.14: Top 90 Australian and English batsmen.

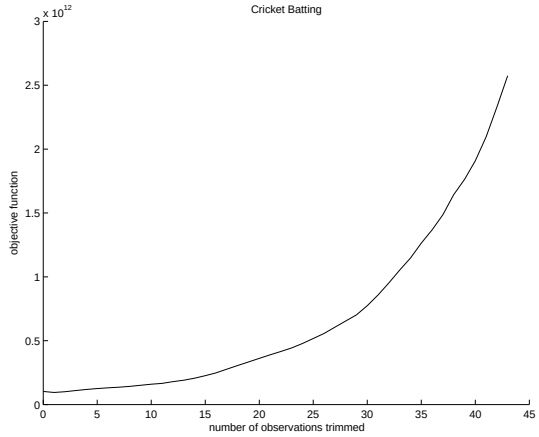


Figure 2.57: Size of (2.7) for subsets chosen by Forward Search.

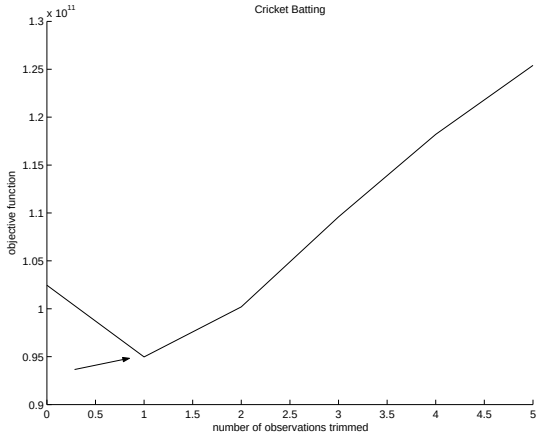


Figure 2.58: Excerpt of Figure 2.57 confirming minimum when Bradman's figures expelled.

a 1 in Figures 2.59-2.60 which suggest this observation is suspiciously outlying. For the second example we choose fifties, hundreds and runs scored and can see in Figure 2.61 a perspective on the data set whereby Bradman does not seem outlying but Figure 2.62 divulges this observations outlyingness. This indicates the usefulness of this method in detecting outliers, there was no way of eyeball pin-pointing some observations as outliers in a 3-dimensional sense without assessing every possible perspective. The new proposal has detected Bradman without the need for plotting the data. Figures 2.63-2.64 illustrate the definitive drop in the measure for the determinant of the asymptotic variance for the location estimate when Bradman's figures were removed from both these three dimensional data sets respectively.

If we consider, for a 2 dimensional example from this data set, fifties scored vs runs amassed, these two variables represent a fairly skewed data set where Bradman is marked with a 1 in Figure 2.65. The new proposal isolated this observation as the sole outlier which is most encouraging since observations 2 and 3 may also appear outlying to the naked eye but are consistent with the trend defined by the majority data.

Figure 2.66 is another good illustration of the impact on (2.7), the removal of Bradman's figures had.

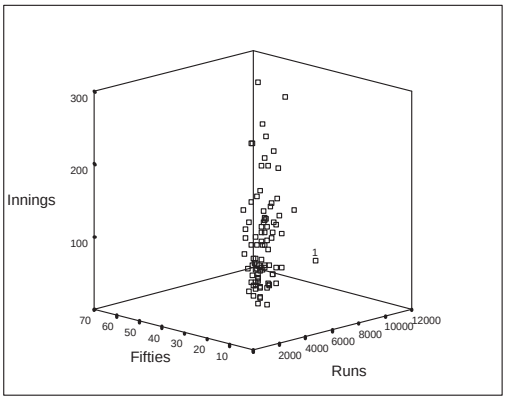


Figure 2.59: Innings, Fifties, Runs (1).

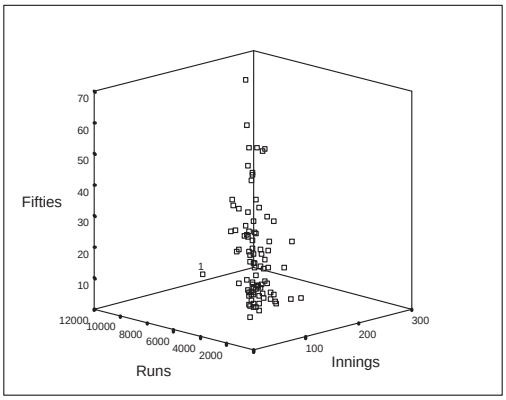


Figure 2.60: Innings, Fifties, Runs (2).

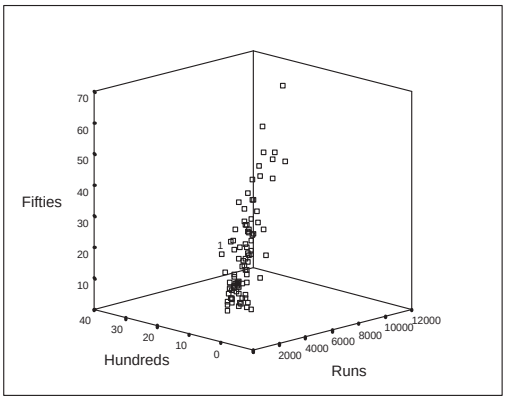


Figure 2.61: Fifties, Hundreds, Runs (1).

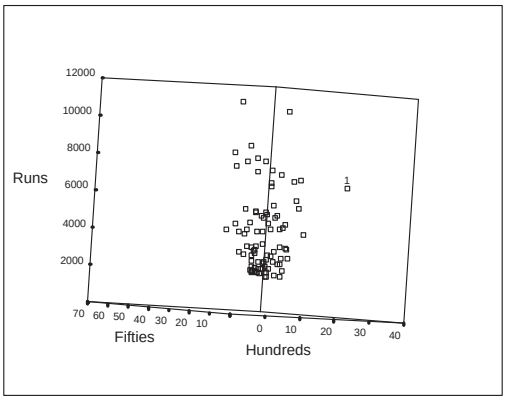


Figure 2.62: Fifties, Hundreds, Runs (2).

2.9 Algorithm for the new Proposal

We are now in a position to describe our proposal step by step. Applying the new proposal **T1** to data sets of size $n \leq 30$ and **T2** otherwise the algorithm is outlined more formally as follows:

Step 1: Calculate robust MCD estimate for location and scale, with $h = \lfloor \frac{n+p+1}{2} \rfloor$.

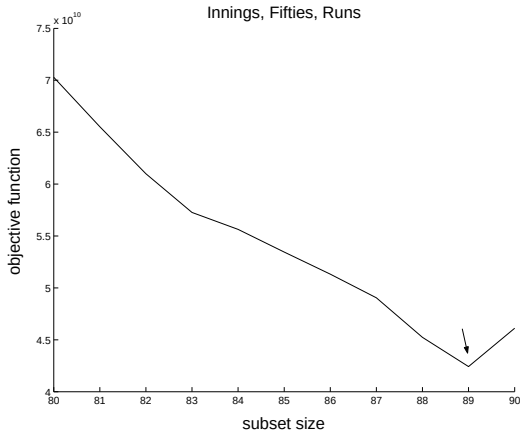


Figure 2.63: Minimum when Bradman expelled.

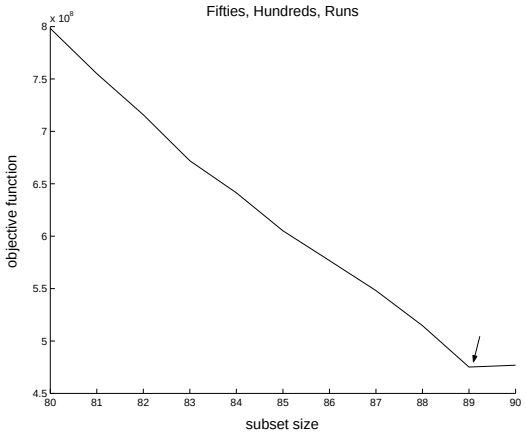


Figure 2.64: Minimum when Bradman expelled.

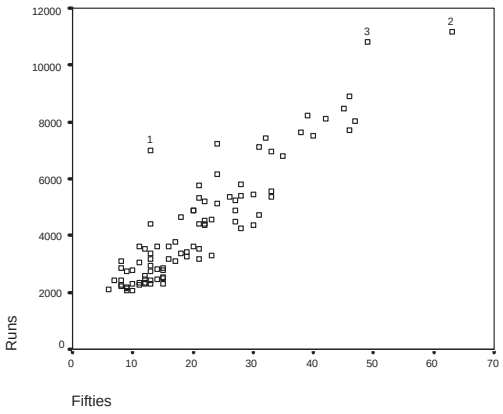


Figure 2.65: Runs vs Fifties

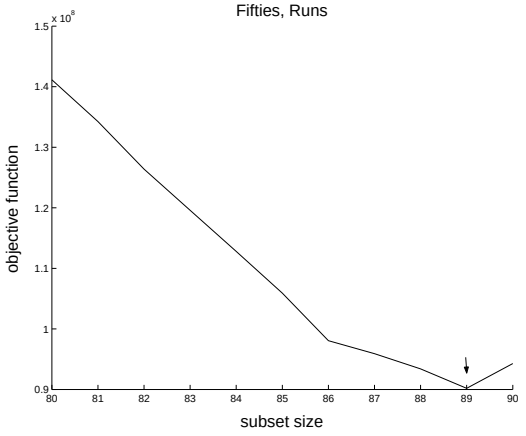


Figure 2.66: (2.7) minimized at $\alpha = 1/90$ when Bradman removed.

Step 2: Order all observations in ascending order of their Mahalanobis distance from this MCD estimate.

Step 3: For the subset of observations retained for the MCD estimate calculate the value of the objective function.

Step 4: Inflate this subset to include that observation within the complement of this subset possessing the smallest Mahalanobis distance from the centroid of this subset.

Step 5: Re-order all observations in ascending order of their Mahalanobis distance with respect to this inflated subset, which may or may not result in the interchange of some members of this retained set and its complement.

Step 6: Assess the value of the objective function for this inflated subset of observations.

Step 7: Repeat steps 4, 5 and 6 until all observations are included in the subset being inflated.

We take that α which corresponds to the minimum of *any* minima occurring for $\alpha > 0$ as the correct trimming proportion. If no minima occur for $\alpha > 0$ then the data set is considered outlier free. If minima occur for an α greater than the minimum minima, $\min_i(\mathsf{m}_i) \;\; i = 1, \ldots, j$, equivalently

$$S_{\gamma_{\min_i(\mathsf{m}_i)}} \neq S_{\gamma_{\mathsf{m}_j}},$$

we can suspect a multi-modal sample, that is we suspect the presence of clusters, and if the sample size is small, one may even dismiss the notion of normality.

Chapter 3

New Robustification of Univariate and Multivariate Regression

3.1 Univariate Regression

In the opening chapter covering robust estimates for multivariate location and scale, various algorithms designed to robustify univariate regression analysis, for example the M-estimate, LMS and LTS were introduced.

The search for a regression estimate with both a high breakdown-point and high efficiency precipitated the introduction of the MM-estimator for regression (Yohai 1987) which is computed using a three step algorithm:

1. Compute a high breakdown-point S-estimator of the regression parameter, that is solve

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \ \hat{s}(\boldsymbol{\beta})$$

such that

$$\frac{1}{n-p}\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\hat{s}(\boldsymbol{\beta})}\right) = K \tag{3.1}$$

where K is generally chosen to be $E_\Phi[\rho]$ where Φ is the standard normal and for maximum breakdown of $\epsilon^* = 1 - \lfloor (n+p+1)/2 \rfloor / n$ is asymptotically 0.5 (Rousseeuw and Yohai 1984,

Lopuhaa 1989).

2. Compute a high breakdown-point M-estimate of the scale parameter by selecting from step 1 the corresponding scale estimate,

$$\hat{s} = \min_{\boldsymbol{\beta}} \ \hat{s}(\boldsymbol{\beta}).$$

3. Calculate an M-estimate which is tuned to have high efficiency (Bianco, Ben and Yohai 2003) using the scale estimate derived in step 2, that is solve for $\hat{\boldsymbol{\beta}}$:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{s}}\right)\mathbf{x}_i = 0 \tag{3.2}$$

where $\psi = \rho'$ is a redescending function.

As an example of the ρ-function in (3.1), the one popularly used for MM-estimates of regression is

$$\rho(x) = \left\{ \begin{array}{ll} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & |x| \leq c \\ \frac{c^2}{6} & |x| \geq c \end{array} \right\},$$

the derivative of which results in Tukey's bi-squared redescending ψ-function (Beaton and Tukey 1974)

$$\psi(x) = \left\{ \begin{array}{ll} x(1 - \frac{x^2}{c^2})^2 & |x| \leq c \\ 0 & |x| \geq c \end{array} \right\}.$$

3.1.1 MMATLA

In the three step algorithm described above, steps 1 and 2 compute highly robust initial estimates for $\hat{\boldsymbol{\beta}}$ and $\hat{s}$ whence step 3 establishes a final estimate for $\hat{\boldsymbol{\beta}}$ by locating the local minimum of (3.2) closest to the initial estimate.

Due to its high efficiency and high breakdown-point the MM-estimate for regression was used to generate a robust fit for 12 data sets from Rousseeuw and Leroy (1987). The subsequent residuals were ordered and subject to the backward deletion method (ATLA),

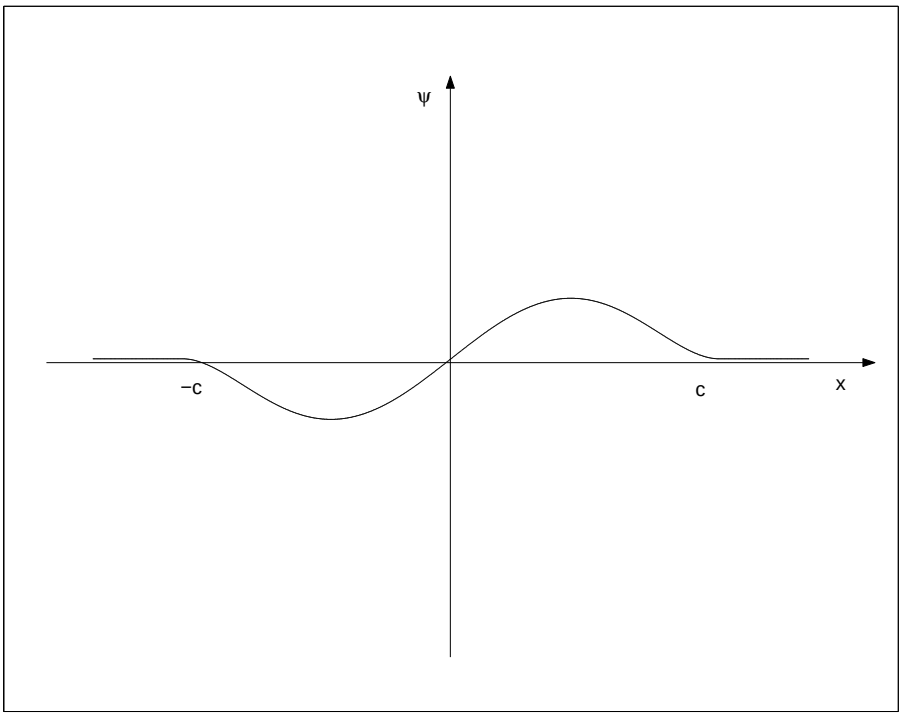


Figure 3.1: Tukey psi function

as explained in Clarke (1994), whereby the residuals, $r_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$, are sorted into ascending order $(r^2)_{1:n} \leq (r^2)_{2:n} \ldots \leq (r^2)_{n:n}$ and the observation associated with the largest remaining residual is deleted. The procedure identifies those observations as outliers which are deleted while minimizing the objective function $V(\alpha, F_n)$ where,

$$V(\alpha, F_n) = \frac{(1-\alpha)\overline{\sigma}_\alpha^2[F_n]}{\{1-\alpha-\sqrt{\frac{2}{\pi}}z_{\alpha/2}e^{(-\frac{z_{\alpha/2}^2}{2})}\}^2} \tag{3.3}$$

and

$$\overline{\sigma}_\alpha^2 = \frac{1}{h-p}\sum_{i=1}^{h}(r^2(T_{MM}))_{i:n}$$

for T_{MM} the robust MM-estimate for location, h signifying the untrimmed number of observations and $r_i(T_{MM}) = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{T_{MM}}$. This objective function is of course identical to that objective function introduced in section 2.1, equation (2.2), slightly modified to account for the changes in degrees of freedom associated with p-dimensional regression. ATLA, in conjunction with the MM-estimate for regression, (MMATLA), was applied to the data sets and Table 3.1 contains the comparison with the results found in Rousseeuw

and Leroy (1987).

Data Set	Known Outliers	Observations Trimmed by ATLA
Wood specific gravity	4,6,8,19	4,6,8,19
Stackloss	1,3,4,21	1,3,4,21
Number of telephone calls	14→ 21	14→ 21
Hawkins-Bradu-Kass	1 →10	1→ 9
Salinity	5,16	16
Coleman	3,17,18	3,18
Pilot-plant	nil	nil
Pilot-plant corrupted	6	6
Hertzsprung-Russell diagram	11,20,30,34	7,9,11,20,30,34
Body and brain weight	6,14,16,17,25	6,7,14,15,16,24,25
Cloud point	1,10,16	1,10,16
Education expenditure	50	50

Table 3.1: Comparison of MMATLA results with Rousseeuw and Leroy (1987).

Table 3.1 indicates the outliers detected by MMATLA are generally consistent with those that have been isolated as legitimate outliers.

For a comparison between robust fits, in conjunction with an application of ATLA, identical tests were carried out using LMS and LTS estimates for regression as a starting point. The results were not as good using LMS and even worse when utilizing an LTS-estimate for robust fit. It is worth noting that Rousseeuw and Leroy (1987) used the LMS-estimate to good effect when locating outliers suggesting MMATLA's increase in complexity is perhaps unnecessary. It should also be noted that the original design of the Adaptive Trimmed Likelihood Algorithm for Regression, (Clarke 2000), involved a computer intensive algorithm to test all possible subsets using backward deletion. This approach achieved more power than this computationally fast counterpart MMATLA, but it's high computational expense renders it useful for only very small data sets.

Monte Carlo experiments were run to get an overall picture of the ability of MMATLA approach to identifying outliers when dealing with robust linear regression models. The data sets are generated from the model for simple regression (Hadi and Simonoff 1993, Clarke 2000),

$$y_i = \beta_0 + \beta_1 x_i + d + \varepsilon_i \qquad i = 1, \ldots, n$$

and for multiple regression,

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + d + \varepsilon_i \quad i = 1, \ldots, n$$

where $x_i \sim U(0, 15)$, $\varepsilon_i \sim N(0, 1)$, $\beta_0 = 0$, $\beta_1 = \beta_2 = 1$ with respect to samples of size $n = 20, 50, 100$. For each sample a pre-specified proportion, $\epsilon = 0, 1/n, 0.1$ respectively, of the sample was displaced, $d = 4$ and $d = 8$ respectively, within neighbourhoods of ± 1 about two positions, $x_i = 7.5$ for low leverage outliers and $x_i = 20$ for high leverage outliers.

Table 3.2 contains the results of the average proportion, $\bar{\alpha} = \overline{(1 - \gamma)}$, of outliers detected by MMATLA which should be close to ϵ, the proportion of the sample displaced. For the larger mean displacement of the outliers the results are excellent independent of outlier positioning, when there are no outliers present in the samples, again the results are excellent with very low proportion of observations being identified as outlying. When the mean displacement of outliers is small, $d = 4$, the average proportion of outliers detected is much lower than those present. For the samples corrupted by a cluster, $\epsilon = 0.1$, of High Leverage outliers the outliers are *rarely* detected. It can be argued that such a small displacement mean for outliers necessarily distributes many contaminants too close to the main sample to be considered outlying.

3.1.2 MMATLA comparison with other robust strategies

Figures 3.2-3.5 compare the MMATLA algorithm with three other outlier detection methods commonly used for univariate regression. The results using these methods, shown labelled MM, LMS and LTS in the figures, were obtained when an observation is identified as an outlier only when it's residual r_i from an MM-fit, LMS-fit or LTS-fit respectively is > 2.5 standardized residuals, s_e, or median absolute deviations (MAD) (Huber 1981,

Simple Regression					*Multiple Regression*				
n	ϵ	d	outlier positioning	$\overline{(1-\gamma)}$	n	ϵ	d	outlier positioning	$\overline{(1-\gamma)}$
20	0			0.0026	20	0			0.0045
	0.05	4	LL	0.0322		0.05	4	LL	0.0350
			HL	0.0271				HL	0.0259
		8	LL	0.0576			8	LL	0.0588
			HL	0.0568				HL	0.0574
	0.1	4	LL	0.0295		0.1	4	LL	0.0556
			HL	0.0509				HL	0.0249
		8	LL	0.1069			8	LL	0.1105
			HL	0.1101				HL	0.0991
50	0			0.0003	50	0			0.0003
	0.02	4	LL	0.01		0.02	4	LL	0.0107
			HL	0.0095				HL	0.0098
		8	LL	0.0214			8	LL	0.0216
			HL	0.0214				HL	0.0216
	0.1	4	LL	0.0149		0.1	4	LL	0.0403
			HL	0.0377				HL	0.0067
		8	LL	0.105			8	LL	0.1045
			HL	0.105				HL	0.1035
100	0			0.00005	100	0			0.0001
	0.01	4	LL	0.004		0.01	4	LL	0.0040
			HL	0.0041				HL	0.0041
		8	LL	0.0104			8	LL	0.0105
			HL	0.0104				HL	0.0105
	0.1	4	LL	0.0055		0.1	4	LL	0.0317
			HL	0.0296				HL	0.0011
		8	LL	0.1063			8	LL	0.1050
			HL	0.1062				HL	0.1049

Table 3.2: Results of MMATLA simulations.

Rousseeuw and Leroy 1987, Venables and Ripley 1999)

$$s_e = \frac{|r_i|}{\text{median}_i |r_i|}.$$

Figures 3.2-3.5 illustrate each algorithm's response to four types of data sets, of size $n = 100$, described by a univariate regression. The first two Figures, 3.2-3.3, for simple regression models as defined above, delineate the average proportion of the samples identified as outlying in relation to the average outlier displacement when the outliers are placed in low leverage and high leverage position respectively.

The latter two Figure, 3.4-3.5, show the comparison between these algorithms for univariate multiple regression as defined above, again we have plotted the average outlier proportions for both low and high leverage positioned outliers. For all of these simulations the proportion of the sample displaced was $\epsilon = 0.1$, denoted by the dashed line, and the outlier displacement ranged from $d = 1, \ldots, 10$.

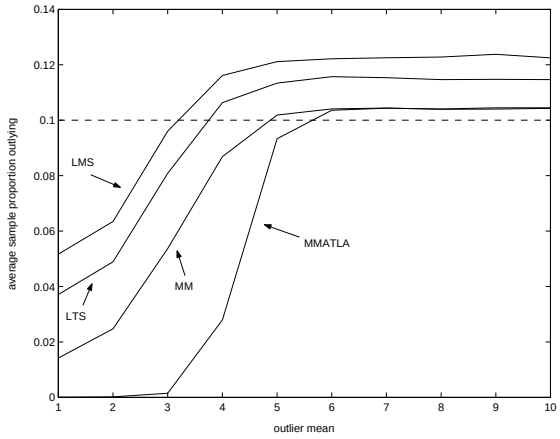


Figure 3.2: Simple Regression Low Leverage.

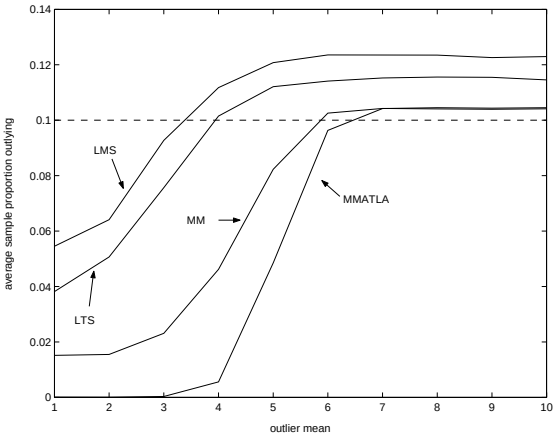


Figure 3.3: Simple Regression High Leverage.

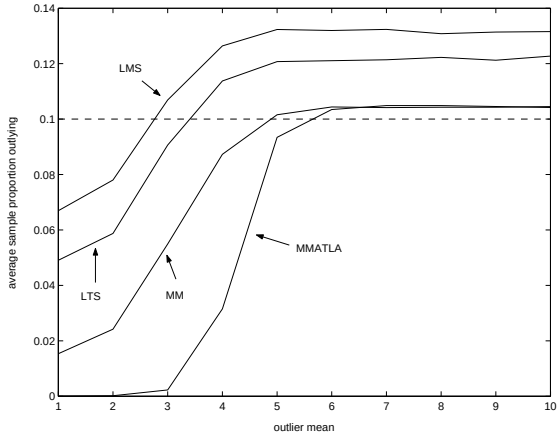


Figure 3.4: Multiple Regression Low Leverage.

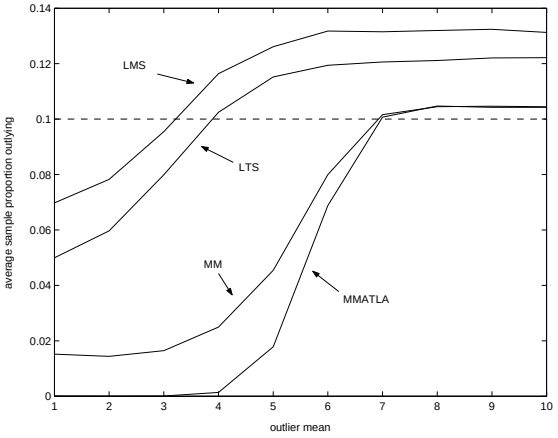


Figure 3.5: Multiple Regression High Leverage.

These figures, 3.2-3.5, show MMATLA is narrowly preferred over the simpler MM-fit whereby outliers are identified if their residuals lie beyond $s_e = 2.5$. The LMS and LTS algorithms appear to be over sensitive to extreme data and although reaching the correct outlier proportion early, overshoot the mark identifying too many observations as outliers. It can be seen for the smallest outlier displacement $d = 1$, MMATLA can be regarded as the more efficient for rarely identifying observations as outliers as it is difficult to imagine these observations truly warranting outlier status.

3.1.3 The new proposal robustifies Univariate Regression

As a preliminary exercise, before using the new proposal to identify outliers corrupting data sets to be described by a multivariate regression, we examine other approaches to univariate data using the **T1** and **T2** proposals.

We transform the data set of observations consisting of predictor variables and *one* response variable into a joint $(\mathbf{x}, y)$ multivariate data set. Whilst in the multivariate phase we apply the new proposal and any outliers detected are removed. Due to its greater sensitivity to outliers than the MM regression fit, an LTS regression was used to establish a robust regression model for the *cleaned* data set and to identify any potential, remaining outliers. We denoted this method **A**:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Apply new proposal and remove any outliers detected.

Step 3: Fit the remaining data set with an LTS regression and remove any residual outliers identified by the LTS.

Method **B** is method **A** with the additional examination of the residuals of the LTS fit using ATLA.

Method **B**:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Apply new proposal and remove any outliers detected.

Step 3: Fit the remaining data set with an LTS regression and remove any residual outliers identified by ATLA.

The third method to be tested against MMATLA, method **C**, is to conduct method **A** whence the cleaned data set is fit by an LTS regression to calculate the regression parameters α, $\boldsymbol{\beta}$. Using these parameters derived from the cleaned data set we proceed to fit *all* the original $n = 100$ points and assess the residuals again using ATLA.

Method **C**:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Apply new proposal and remove any outliers detected.

Step 3: Fit the remaining data set with an LTS regression deriving estimates for regression

n	ϵ	d	outlier positioning	*MMATLA* $\overline{(1-\gamma)}$	*Method* ***A*** $\overline{(1-\gamma)}$	*Method* ***B*** $\overline{(1-\gamma)}$	*Method* ***C*** $\overline{(1-\gamma)}$
20	0			0.0045	0.018	0.0089	0.0161
	0.05	4	LL	0.035	0.0158	0.0.0450	0.0521
			HL	0.0259	0.0256	0.0435	0.0467
		8	LL	0.0588	0.0458	0.0614	0.0803
			HL	0.0574	0.0482	0.0658	0.0791
	0.1	4	LL	0.0556	0.0232	0.0804	0.0711
			HL	0.0249	0.0575	0.0772	0.0739
		8	LL	0.1105	0.0945	0.1167	0.139
			HL	0.0991	0.0993	0.1118	0.1351
50	0			0.0003	< 0.0001	0.0003	0.0003
	0.02	4	LL	0.0107	0.0022	0.0103	0.0106
			HL	0.0098	0.011	0.0124	0.0106
		8	LL	0.0216	0.0197	0.0204	0.0223
			HL	0.0216	0.02	0.0206	0.0221
	0.1	4	LL	0.0446	0.008	0.0435	0.0434
			HL	0.0067	0.0892	0.0368	0.0496
		8	LL	0.1045	0.1003	0.1021	0.1074
			HL	0.1035	0.1001	0.1005	0.1055
100	0			0.0001	< 0.0001	0.0001	0.0001
	0.01	4	LL	0.004	0.0008	0.0043	0.0039
			HL	0.0041	0.0051	0.0054	0.0045
		8	LL	0.0105	0.0099	0.01	0.0104
			HL	0.0105	0.01	0.0101	0.0104
	0.1	4	LL	0.0317	0.0014	0.0381	0.0379
			HL	0.0011	0.097	0.0327	0.0422
		8	LL	0.105	0.1004	0.1018	0.1054
			HL	0.1049	0.1004	0.1005	0.1052

Table 3.3: Simulation results for MMATLA, method **A**, **B** and **C** applied to Multiple Regression models.

parameters.

Step 4: Fit the entire, original data set using the parameters derived in Step 3.

Step 5: Remove any residual outliers detected by ATLA.

Table 3.3 contains the average proportion, $\overline{(1-\gamma)}$, of outliers detected for each of the 4 algorithms when applied to univariate multiple regression models for data sets of size $n = 100$, contaminated with a proportion of $\epsilon = 1/n, 0.1$ displaced observations about an outlier mean of $d = 4$ and $d = 8$ respectively.

Figures 3.6-3.7 present a graphical representation of the comparison between MMATLA

and methods **A, B** and **C** when used for outlier detection on multiple regression models used to describe the Monte Carlo samples generated as before.

With regard to Low Leverage contaminants (see page 102) we see method **B** performing better than the others, method **A** is as accurate for the larger outlier displacements but less sensitive whilst MMATLA and **C** tend to be too sensitive as d increases. For High Leverage outliers method **A** performed as efficiently as method **B** whilst MMATLA and method **C** again appear to consistently identify too many observations as outliers as well as being less sensitive for the smaller outlier displacements d.

Method **B** appears the best overall but one may prefer using the simpler method **A**.

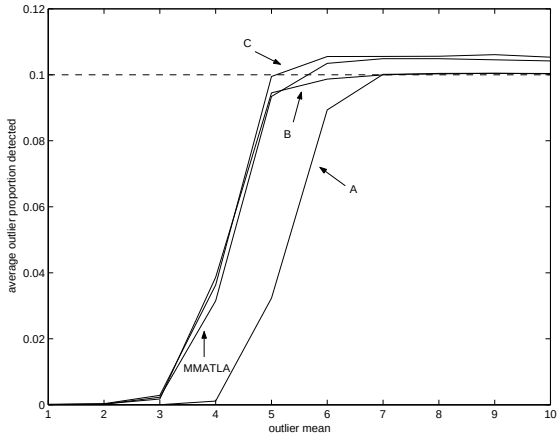


Figure 3.6: Multiple MMR Regression Low Leverage

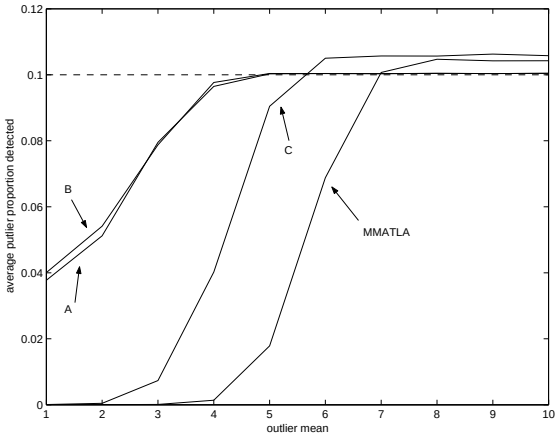


Figure 3.7: Multiple MMR Regression High Leverage

3.1.4 2 real data sets revisited

We revisit two examples from Rousseeuw and Leroy (1987), for which results using MMATLA are documented in Table 3.1, for an application of method **A** to the 4 dimensional Salinity data set and the 6 dimensional Wood Specific Gravity data set. When **T1** is used to assess the joint $(\mathbf{x}, y)$ Salinity data, $n = 28$, again only observation 16 was considered

an outlier although it was noted that when **T2** was applied, observation 5, supposedly masked by observation 3 (Rousseeuw and Leroy 1987), was also identified as outlying. When considered as a joint $(\mathbf{x}, y)$ multivariate sample, the Wood Specific Gravity data had the 4 known outliers, $4, 6, 8$ and 19 detected by **T1**. Figures 3.8-3.9 depict the size of the objective functions when **T2** and **T1** were applied to the Salinity and Wood Specific Gravity data sets respectively, in relation to the number of observations trimmed.

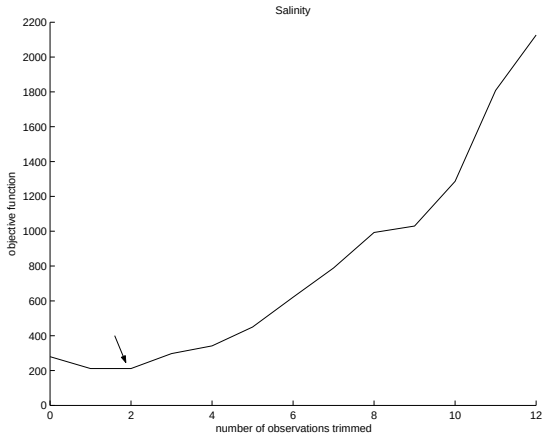


Figure 3.8: Method A on Salinity.

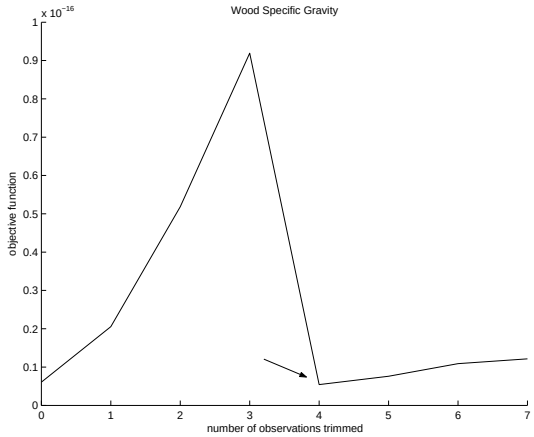


Figure 3.9: Method A on Wood Specific Gravity

3.2 Multivariate Regression

In the case of classical linear regression, one is dealing with a single *response* variable, described by an arbitrary number of *predictor* variables, and an associated random error due to measurement error and other unknown factors and variables. In the multivariate case we have more than one response variable comprising the effects of the same predictors.

$$Y_1 = \beta_{01} + \beta_{11}x_1 + \beta_{21}x_2 + \ldots + \beta_{p1}x_p + \varepsilon_1$$

$$Y_2 = \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2 + \ldots + \beta_{p2}x_p + \varepsilon_2$$

$$\vdots$$

$$Y_q = \beta_{0q} + \beta_{1q}x_1 + \beta_{2q}x_2 + \ldots + \beta_{pq}x_p + \varepsilon_q$$

where $\mathbf{Y}$ represents the response vectors, $\boldsymbol{\beta}$ represents the unknown parameter vectors and $\mathbf{X}$ contains the values of the predictor vectors. With $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ the multivariate linear regression model can be represented by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where, by Johnson and Wichern (1998),

$$\text{Y} = \begin{bmatrix} Y_{11} & Y_{12} & \ldots & Y_{1q} \\ Y_{21} & Y_{22} & \ldots & Y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \ldots & Y_{nq} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \ldots & \beta_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \ldots & \beta_{pq} \end{bmatrix},$$

$$\text{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \ldots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & \ldots & \varepsilon_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \ldots & \varepsilon_{nq} \end{bmatrix}.$$

We can view the known values

$$\mathbf{x}_k^\top = (x_{k1}, \ldots, x_{kp}),$$

$$\mathbf{Y}_k^\top = (Y_{k1}, \ldots, Y_{kq}),$$

for $k = 1, \ldots, n$, as joint (x_k, y_k) variables of length $(p + q)$ to comprise an overall joint $(\mathbf{x}, \mathbf{y})$ multivariate sample. Using the notation of Rousseeuw et al (2004),

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_{\mathbf{x}} \\ \hat{\boldsymbol{\mu}}_{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{Y}} \end{pmatrix}$$

and

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{xy}} \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{yx}} & \hat{\boldsymbol{\Sigma}}_{\mathbf{yy}} \end{pmatrix},$$

we can calculate the least squares estimators for the slope matrix $\hat{\boldsymbol{\beta}}_{ij} = \hat{\boldsymbol{\beta}},\ i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, q$, the multidimensional intercept $\hat{\boldsymbol{\beta}}_{0j} = \hat{\boldsymbol{\alpha}}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\mathcal{E}}}$ using (Rousseeuw et al 2004),

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{xy}},$$

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_{\mathbf{y}} - \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\mu}}_{\mathbf{x}}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\mathcal{E}}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{yy}} - \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}\hat{\boldsymbol{\beta}}.$$

3.2.1 Robust Multivariate Regression Algorithms

Wisnowski et al (2002) had success using a modified Simpson and Montgomery (1998) and Coakley and Hettmansperger (1993) Compound Estimators. Each of these algorithms were quite extensive, beginning with an initial estimate of $\boldsymbol{\beta}$ using the already discussed LTS or S-estimators. Using this estimate they seek to minimize

$$\min_{\beta} \sum_{i=1}^{n} \pi_i \frac{\rho(y_i - \mathbf{x}_i^{\top}\hat{\boldsymbol{\beta}})}{s\pi_i} \tag{3.4}$$

where π_i is a measure of leverage. This measure corresponds to the Minimum Volume Ellipsoid robust distances, Coakley and Hettmansperger (1993), or an M-estimate of covariance, Simpson and Montgomery (1998), i.e.

$$\text{diag}\ \ \mathbf{x_i}(\mathbf{X}^{\top}\mathbf{X})^{-\mathbf{1}}\mathbf{x_i}^{\top}.$$

The solution to (3.4) is most commonly approached using Newtons method, Coakley and Hettmansperger (1993), or an iterative reweighted least squares, Simpson and Montgomery (1998).

The modification proposed by Wisnowski et al (2002) details an improvement to the initial estimate of $\boldsymbol{\beta}$. This improvement involves a 3 stage algorithm to compute the initial estimate. At the first stage they filter out any high leverage points using the fixed threshold procedure of Rocke and Woodruff (1996). Step 2 is designed to filter out any

residual outliers using an MM-estimator for regression, Yohai (1987). The ensuing MM-estimate for regression may be used to fit the remaining observations or an Ordinary Least Squares fit is used. After the initial estimate has been computed, Wisnowski et al (2002), then proceed as is the case with Coakley and Hettmansperger (1993) and Simpson and Montgomery (1998).

More recently, Rousseeuw et al (2004) attempted 3 different weighting methods to improve the efficiency of their estimates. Initially they treat the p predictor variables, $\mathbf{x}$, and q response variables, $\mathbf{y}$, as the joint $(\mathbf{x}, \mathbf{y})$ variables of a multivariate data set in and of itself, then identify outliers as those points, in the multidimensional space, who's Mahalanobis distance from an MCD estimate for location, is beyond a pre-specified cutoff, for example, $M_{\text{outlier}} > \sqrt{\chi^2_{0.99,p+q}}$. Even better results were obtained using a reweighted regression algorithm whereby the residuals were assessed for outlying information which ignores possible outlier's if they are good leverage points. The methodology most successful was to combine theses two methods, assess the joint $(\mathbf{x}, \mathbf{y})$ for outlying data, using that data considered inlying, construct a regression based on the retained data and use the resulting robust $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to calculate the residuals for the *original* data set in its *entirety*. The final step is to inspect these residuals for any outlying data, this will again prevent the identification of good leverage points as outliers.

3.2.2 Simulation models

For simulations involving multivariate multiple regression we investigated two model types applied to sample sizes $n = 20, 50, 100$ respectively. The model types concerned $p = 2$ predictor variables with $q = 2$ response variables and $p = 4$ predictor variables with $q = 4$ response variables. Each variable was composed of a randomly generated data set $\sim N(0, 1)$ and each sample consisted of one of three levels of contamination:

- **CL1**: *no* outliers planted, corresponding to $\epsilon = 0$.
- **CL2**: A proportion, $\epsilon_1 = 0.1 - 1/n$, of *vertical* outliers and a proportion, $\epsilon_2 = 1/n$,

of *bad leverage* outliers, for example concerning sample size of $n = 100$, nine of the q response variables were distributed $N(2\sqrt{\chi^2_{0.99,p+q}}, 0.1)$ parading as vertical outliers and one response variable is distributed $N(2\sqrt{\chi^2_{0.99,p}}, 0.1)$ along with its corresponding p predictor variables, a bad leverage, high impact outlier. Note $p = q$ for these simulations.

• **CL3**: A proportion, $\epsilon_1 = 0.1$, of the sample was planted with vertical outliers and a proportion, $\epsilon_2 = 0.1$, of the sample was planted with bad leverage points all distributed consistent with the contamination for the second level.

Figure 3.10 provides a graphical representation of the outlier configurations for the three levels imposed. For each level we can see the *robust* distances of the observations with respect to the robust distances of the residuals after an LTS fit on the data without any trimming. The distances are robust in the sense that they are in fact Mahalanobis distances from an MCD estimate for location and scale for the observations and the residuals respectively. We can see from these example data sets that even for outlier free samples there will be extreme observations lying beyond the $\sqrt{\chi^2_{0.975,p+q}} = 3.34$ value denoted by the dashed lines.

3.2.3 New proposals for Multivariate Regression

There will be two new methodologies assessed, the first method **R1** is where the joint $(\mathbf{x}, \mathbf{y})$ variables are treated as a $p+q$ dimensional data set and to which the new proposal, **T1** or **T2** depending on sample size, will be applied. The data set is then stripped of any outliers identified by this new proposal and the remaining data points are modelled by LTS regression. The second method to be assessed, method **R2**, is based on Rousseeuw et al (2004) whereby the resulting parameter estimates, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, derived from the LTS analysis of the initial retained subset of data points are used to calculate the residuals for all the original data points. These residuals are assessed for outliers again by using the new proposal, **T1** or **T2**, and any points identified as outliers at this stage remain as outliers. The advantage with the new proposal is we do not need to specify a cut-off

region because the new proposal determines this for us.

R1:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Apply new proposal removing any outliers detected.

Step 3: Model remaining points using an LTS regression.

R2:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Apply new proposal removing any outliers detected.

Step 3: Model remaining points using an LTS regression.

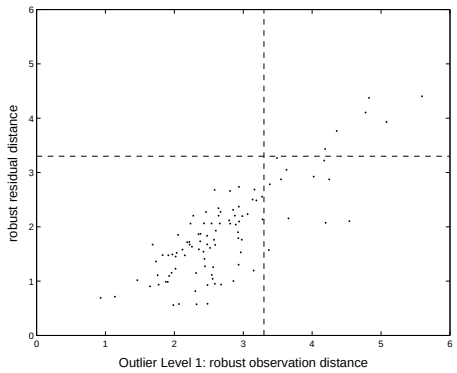
Step 4: Fit entire, original data set using the regression parameters derived in Step 3.

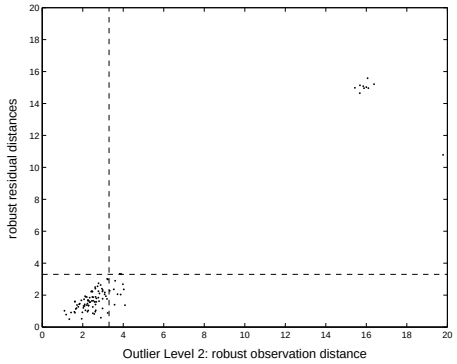
Step 5: Apply new proposal to identify any residual outliers.

Table 3.4 depicts $\overline{(1-\gamma)}$, the average proportion of outliers detected, when using **T1** for samples sizes of $n = 20$ and **T2** otherwise, when **R1** and **R2** were applied to data sets compose of $p = 4$ predictor variables and $q = 4$ response variables. The figures are strong for the sample sizes $n = 50, 100$, recalling that scenario **CL1** represents clean data sets, **CL2** correspond to a total corrupted proportion of $\epsilon_1 + \epsilon_2 = 0.1$ and **CL3** an outlying proportion of $\epsilon_1 + \epsilon_2 = 0.2$ overall. For samples of size $n = 20$ there are definitely too many observations being identified as outlying on average. As already noted, the latter shortcoming is not particularly dangerous to statistical inference since trimming good observations, along with the bad, will reduce estimate efficiency but not greatly impact the values of parameter estimates. Any outliers *not* detected would surely corrupt any parameter estimates. The most important aspect covered in Table 3.4 is the negligible difference between using **R1** and the more complicated **R2**.

n	method	ϵ_{TOTAL}	$\overline{(1-\gamma)}$	n	method	ϵ_{TOTAL}	$\overline{(1-\gamma)}$	n	method	ϵ_{TOTAL}	$\overline{(1-\gamma)}$
20	**R1**	0	0.078	50	**R1**	0	0.0001	100	**R1**	0	< 0.0001
		0.1	0.1026			0.1	0.1011			0.1	0.1003
		0.2	0.1815			0.2	0.2014			0.2	0.2010
	R2	0	0.1183		**R2**	0	0.0012		**R2**	0	< 0.0001
		0.1	0.1244			0.1	0.1051			0.1	0.1004
		0.2	0.2356			0.2	0.2051			0.2	0.2006

Table 3.4: Outlier detection accuracy using **R1** and **R2**, $p = q = 4$.





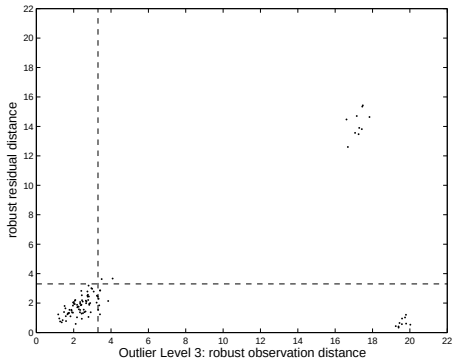


Figure 3.10: Diagnostic plots for three contamination levels.

Figures 3.11-3.12 represent a comparison between **R1** and **R2** with two other algorithms based on a generalized fixed-threshold methodology. **R3** involves using the approach of **R1** without the application of the new proposal to locate outliers, instead when using **R3**, prior to LTS modelling the data is cleaned of outliers identified as such if they lie beyond $\sqrt{\chi^2_{0.99,p+q}}$ from the MCD estimate for location (Rousseeuw et al 2004). **R4** follows the same procedure as **R2** with the application of this fixed cut-off value to the residuals of the LTS fit.

R3:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Identify as outliers any point with a Mahalanobis distance $M > \sqrt{\chi^2_{0.99,p+q}}$ from an MCD estimate for location.

Step 3: Model remaining points using an LTS regression.

R4:

Step 1: Transform data set into a joint $(\mathbf{x}, y)$ data set.

Step 2: Apply new proposal removing any outliers detected.

Step 3: Model remaining points using an LTS regression.

Step 4: Fit entire, original data set using the regression parameters derived in Step 3.

Step 5: Identify as outliers any residual with a value greater than $\sqrt{\chi^2_{0.99,p+q}}$.

Figures 3.11-3.12 were derived from simulations using these methodologies applied to data sets consisting of $p = 2$ predictor variables and $q = 2$ response variables and compare the average proportion of outliers detected by these methods with reference to the true proportion planted, ϵ, denoted by the dashed line. Notice the range of outlier means, $d = 0, \ldots, 10$, for both the outlying scenarios **CL2** and **CL3**, described above, *also* yields the proportion of outliers detected, $\overline{(1-\gamma)}$, when the data sets are in fact clean, $d = 0$, which corresponds to outlier level **CL1**. $\overline{(1-\gamma)}$ ideally should be zero at this point. The plots, Figures 3.11-3.12, reveal the only real difference between the 4 methods is the slight tendency to identify outliers in clean data sets by methods **R3** and **R4**.

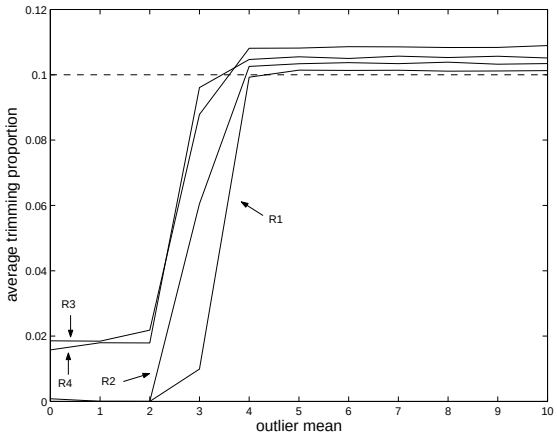


Figure 3.11: Outlier Level **CL2**.

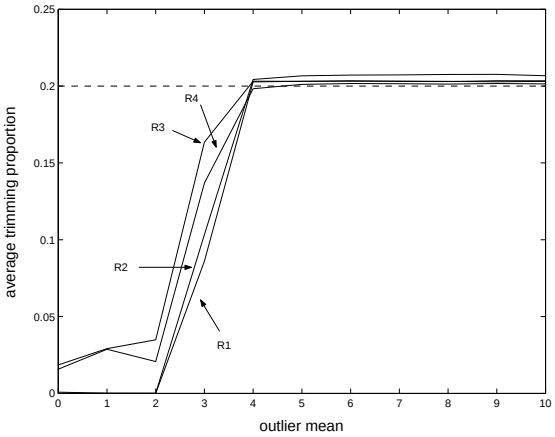


Figure 3.12: Outlier Level **CL3**.

3.2.4 Bias and MSE tests

Further assessment of the algorithms, **R1** and **R2**, when applied to data sets with a pre-specified level of outlying data for various sample sizes involves calculating the ensuing bias and Mean Squared Error of the slope matrix, intercept and error matrix (Rouseeuw et al 2004),

$$\text{bias}(\hat{\boldsymbol{\beta}}) = \sqrt{\text{ave}_{ij}(\text{bias}(\hat{\beta}_{ij})^2)} \quad i = 1, \ldots, p,\ j = 1, \ldots, q,$$

and

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{ave}_j(\text{MSE}(\hat{\beta}_{ij})).$$

where $\text{bias}(\hat{\beta}_{ij})^2$ will simply be $\hat{\beta}_{ij}^2$ since the randomly generated data sets will ideally set $\beta_{ij} = 0$, as will be the case for the intercept, $\boldsymbol{\alpha} = 0$. The estimated error matrix will be compared with the ideal $\boldsymbol{I}_p$. Obviously the bias and MSE should be as small as possible for any methodology to be useful.

Tables 3.5-3.6 contain the Monte Carlo results for Bias and MSE for the methods **R1** and **R2** for the $p = q = 4$ setting. Table 3.7 contains the results for the statistics assessed in

Tables 3.5-3.6 for clean data sets when *no* trimming algorithm was applied, such figures should reflect the best results one can hope to achieve using any methodology. We can observe from Tables 3.5-3.6 in comparison with Table 3.7 that method **R1** has performed better than method **R2** for both clean and contaminated data sets. Note this method **R1** is an extension the method decided upon when dealing with univariate data sets cast in a multivariate setting, method **A**.

Outlier Level	parameter	*Bias*			*MSE*		
		$n = 20$	$n = 50$	$n = 100$	$n = 20$	$n = 50$	$n = 100$
CL1	Slope	0.0153	0.0073	0.0061	6.2932	2.1728	1.7378
	Intercept	0.0131	0.0031	0.0022	4.4464	1.8617	1.6656
	$\Sigma_{offdiag}$	0.0117	0.0047	0.0022	1.3845	1.0284	0.9892
	Σ_{ondiag}	0.2091	0.0158	0.016	12.0801	2.6032	2.1803
CL2	Slope	0.0135	0.0092	0.0031	4.4146	2.59	1.9457
	Intercept	0.0182	0.0033	0.0017	3.4939	2.186	1.7809
	$\Sigma_{offdiag}$	0.0133	0.0029	0.0036	1.4063	1.1505	1.0857
	Σ_{ondiag}	0.0448	0.0104	0.0179	11.0429	3.1399	2.3872
CL3	Slope	0.0147	0.0067	0.0044	5.8516	2.9488	2.3060
	Intercept	0.0119	0.0058	0.0052	4.0856	2.5273	2.0708
	$\Sigma_{offdiag}$	0.0522	0.0048	0.0029	11.5691	1.219	1.2389
	Σ_{ondiag}	0.0151	0.0206	0.0202	52.4948	3.9006	2.8562

Table 3.5: Method ***R1*** p=4, q=4.

Outlier Level	parameter	*Bias*			*MSE*		
		$n = 20$	$n = 50$	$n = 100$	$n = 20$	$n = 50$	$n = 100$
CL1	Slope	0.0205	0.0063	0.0049	7.3301	2.1948	1.7558
	Intercept	0.0046	0.0053	0.0031	5.6009	1.8743	1.6619
	$\Sigma_{offdiag}$	0.0063	0.0036	0.0027	0.9885	0.9866	1.0028
	Σ_{ondiag}	0.34	0.0164	0.0124	16.197	2.8909	2.1744
CL2	Slope	0.0132	0.0083	0.0059	5.1683	2.5812	1.9968
	Intercept	0.008	0.006	0.0035	3.9016	2.2077	1.8317
	$\Sigma_{offdiag}$	0.0081	0.0025	0.0038	1.1169	1.1487	1.1302
	Σ_{ondiag}	0.18	0.0287	0.0254	11.5623	3.2406	2.5220
CL3	Slope	0.0203	0.0092	0.0041	7.0833	3.0748	2.3014
	Intercept	0.0162	0.0066	0.0058	5.1958	2.6557	2.1659
	$\Sigma_{offdiag}$	0.0049	0.0049	0.0032	1.1463	1.295	1.2680
	Σ_{ondiag}	0.2585	0.0448	0.0328	19.1642	4.2303	3.0615

Table 3.6: Method ***R2*** p=4, q=4.

parameter	*Bias*			*MSE*		
	$n = 20$	$n = 50$	$n = 100$	$n = 20$	$n = 50$	$n = 100$
Slope	0.0115	0.0062	0.0048	3.0085	2.1182	1.7457
Intercept	0.0066	0.0072	0.0055	2.4393	1.8745	1.6288
$\Sigma_{offdiag}$	0.0096	0.003	0.0044	1.1577	1.042	0.9684
Σ_{ondiag}	0.0311	0.0089	0.0173	6.7183	2.6519	2.2238

Table 3.7: Clean data, no trimming algorithm applied p=4, q=4.

Figures 3.13-3.16 portray a comparison of the corresponding slope and intercept MSE results for the 4 methodologies, **R1**, **R2**, **R3** and **R4**, when applied to data sets consisting of $p = 2$ predictor variables and $q = 2$ response variables. Again as was the case illustrated in Figures 3.11-3.12, we have a measure of the MSE's over a range of different outlier means, $d = 0, \ldots, 10$, for outlier levels 1 and 2. The performance of the methodologies, when applied to samples contaminated to these levels, is explicitly shown whilst the figures associated with outlier level, **CL1**, are shown for when the outlier mean is zero, $d = 0$.

These Figures, 3.13-3.16, show that all four methodologies, **R1**, **R2**, **R3** and **R4** respectively, perform to much the same degree of success. When one compares the results for outlier means of $d \geq 4$ there is a levelling off of the MSE's very comparable to those when the samples were not corrupted with outliers, level **CL1** contamination, after a spike in the MSE values for all four methods for outlier means on the interval $1 < d < 4$.

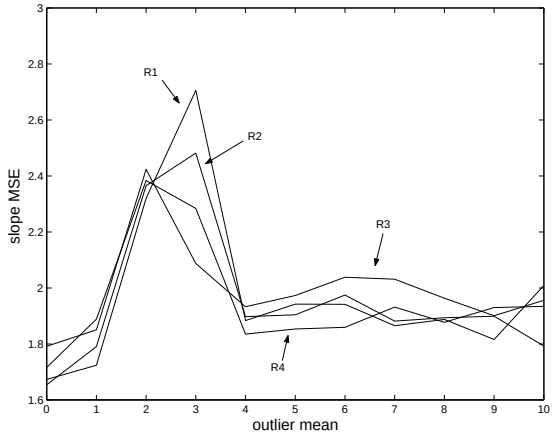


Figure 3.13: Slope MSE **CL2**.

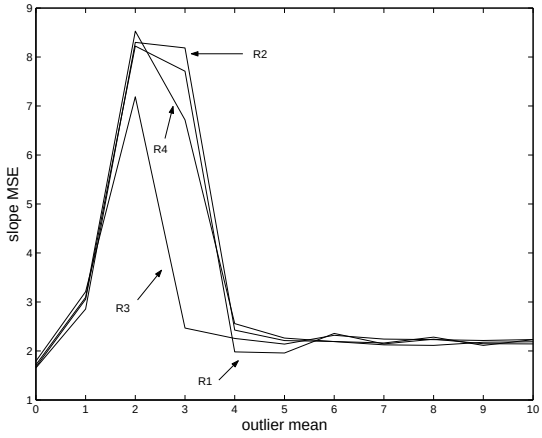


Figure 3.14: Slope MSE **CL3**.

3.2.5 Finite-Sample Efficiencies

Further exploration into the success of methods **R1**, **R2** is to construct Finite-Sample Efficiencies (Rousseeuw et al 2004). The finite-sample efficiencies are construed via

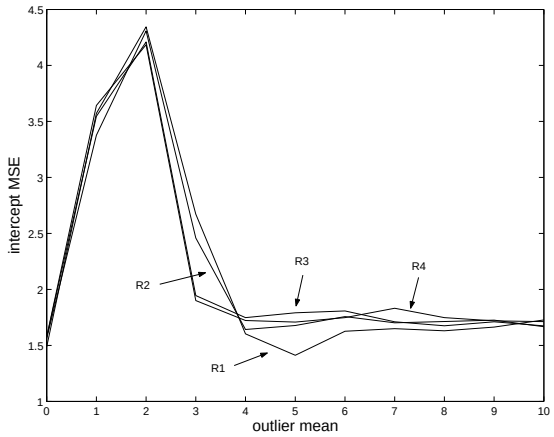


Figure 3.15: Intercept MSE **CL2**.

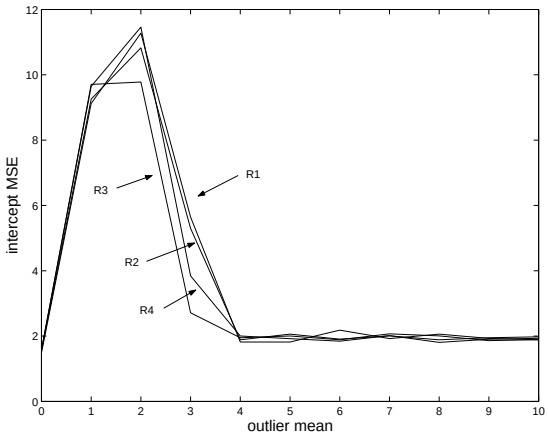


Figure 3.16: Intercept MSE **CL3**.

$$\text{var}(\hat{\boldsymbol{\beta}}) = \text{ave}_{ij}(\text{var}(\hat{\boldsymbol{\beta}}_{ij}))$$

where

$$\text{var}(\hat{\boldsymbol{\beta}}_{ij}) = n \ \text{var}_l(\hat{\boldsymbol{\beta}}_{ij}^{(l)})$$

for sample size n over $l = 1, \ldots, N$ simulations.

The corresponding finite-sample efficiency is given by $\frac{1}{(\text{var}(\hat{\boldsymbol{\beta}}_{ij}))}$ and similarly $\frac{1}{(\text{var}(\hat{\boldsymbol{\alpha}}_j))}$. For assessment of the resulting error covariance matrix $\hat{\boldsymbol{\Sigma}}_\epsilon$ we define (Bickel and Lehmann 1976, Rouseeuw et al 2004) a standardized variance

$$\text{StdVar}((\hat{\boldsymbol{\Sigma}}_\epsilon)_{ij}) = \frac{n\text{var}_l((\hat{\boldsymbol{\Sigma}}_\epsilon^{(l)})_{ij})}{[\text{ave}_l\text{ave}_j((\hat{\boldsymbol{\Sigma}}_\epsilon^{(l)})_{jj})]^2}$$

whence the overall finite-sample efficiencies of the error covariance matrix may be computed as

$$\frac{1}{\text{ave}_{i\neq j}(\text{StdVar}((\hat{\boldsymbol{\Sigma}}_\epsilon))_{ij})},$$

for the off diagonal elements and

$$\frac{2}{\text{ave}_j(\text{StdVar}((\hat{\boldsymbol{\Sigma}}_\epsilon))_{jj})},$$

Outlier level	Finite-Sample efficiencies	$n = 20$	$n = 50$	$n = 100$
1	Slope	0.1589	0.4603	0.5761
	Intercept	0.2249	0.5367	0.6000
	$\Sigma_{offdiag}$	0.453	0.9457	0.9785
	Σ_{ondiag}	0.1117	0.75	0.8981
2	Slope	0.2265	0.3863	0.5137
	Intercept	0.2865	0.4571	0.5610
	$\Sigma_{offdiag}$	0.6514	0.8579	0.8886
	Σ_{ondiag}	0.1661	0.6296	0.8183
3	Slope	0.1708	0.339	0.4336
	Intercept	0.2446	0.3956	0.4831
	$\Sigma_{offdiag}$	0.0871	0.7907	0.7761
	Σ_{ondiag}	0.0382	0.4964	0.6826

Table 3.8: Method **R1** p=4, q=4.

Outlier level	Finite-Sample efficiencies	$n = 20$	$n = 50$	$n = 100$
1	Slope	0.1364	0.4555	0.5698
	Intercept	0.1784	0.5333	0.6014
	$\Sigma_{offdiag}$	0.4411	0.9848	0.9727
	Σ_{ondiag}	0.0628	0.6749	0.9030
2	Slope	0.1934	0.3875	0.5012
	Intercept	0.2561	0.4529	0.5458
	$\Sigma_{offdiag}$	0.6029	0.822	0.8416
	Σ_{ondiag}	0.1232	0.5901	0.7732
3	Slope	0.1412	0.3253	0.4344
	Intercept	0.1924	0.3765	0.4620
	$\Sigma_{offdiag}$	0.4802	0.7061	0.7379
	Σ_{ondiag}	0.0617	0.4424	0.6630

Table 3.9: Method **R2** p=4, q=4.

Finite-Sample efficiencies	$n = 20$	$n = 50$	$n = 100$
Slope	0.3324	0.4721	0.5730
Intercept	0.4097	0.5337	0.6145
$\Sigma_{offdiag}$	0.9068	0.9437	0.9985
Σ_{ondiag}	0.3129	0.7424	0.8798

Table 3.10: Clean data sets, no trimming algorithm imposed, p=4, q=4.

for the diagonal elements (Rouseeuw et al 2004).

Tables 3.8-3.9 contain the Finite-Sample Efficiencies relating to the method ***R*1** applied to data sets incurring each of the 3 levels of contamination for the $p = 4,\ q = 4$ setting. A comparison with those finite-sample efficiencies reported in Table 3.10, where by the finite sample efficiencies were calculated for clean data sets not subject to any trimming algorithm, show that method **R1** is again the preferred method when detecting outliers whilst applying a multivariate regression analysis.

3.3 Regression with Correlated Variables

For the sake of completeness it was decided to apply **R1** to a simulation model involving two correlated variables for sample sizes of $n = 100$. The contamination types were **CL2** and **CL3** and the model was again composed of $p = 4$ predictor variables and $q = 4$ response variables. The simulation regression models were as described in Section 3.2.2 only for these trials the 3rd predictor variable, say x_3, was correlated with the the 2nd predictor variable x_2, so that

$$x_3 = 10x_2 + \varepsilon$$

where $\varepsilon \sim N(0, 1)$. The results, shown in Table 3.11, were again promising, reflecting the same healthy trend as found for the non-correlated regression models.

	CL1			**CL2**		
Parameter	*Bias*	*MSE*	*Finte Sample Efficiency*	*Bias*	*MSE*	*Finte Sample Efficiency*
slope	0.0197	1.2570	0.3900	0.1047	3.3193	0.3180
Intercept	0.0021	1.7128	0.5834	0.0320	2.0919	0.4976
$\Sigma_{offdiag}$	0.0029	1.0887	0.8854	0.0103	1.3163	0.7312
Σ_{ondiag}	0.0196	2.4763	0.7902	0.0248	2.6207	0.7461

Table 3.11: $n = 100$, $p = 4$, $q = 4$, Regression with Correlated Variables.

The corresponding average trimming proportions were also very good, $\overline{(1-\gamma)}_{\mathbf{CL1}} = 0.1004$ and $\overline{(1-\gamma)}_{\mathbf{CL2}} = 0.2014$ respectively.

Chapter 4

Using an Adaptive Trimmed Likelihood for Cluster Detection

Towards the end of Chapter 2 we discussed the minima selection strategy when confronted with multiple minima, in this chapter we expand on Schubert (2006a) probing further into specific types of contamination which yield multiple minima in the objective function. In most cases, to reiterate, when multiple minima occured for an $\alpha > 0$, the *minimum* of these minima corresponded with the *minimum subset* of retained data. Certain specific *outlying* scenarios warrant closer attention because there were cases when the minimum of the minima did *not* correspond with the smallest subset of retained data, or equivalently the greatest value of the corresponding trimming proportion α. An inspection of the ensuing minima for such data sets divulged clustered configurations composing the sample had been detected.

Recapitulating, the new algorithm involves a preliminary robust minimum covariance determinant, MCD, estimate for location and scale followed by the the use of a Forward Search algorithm. For each subset of the p-dimensional data selected by the Forward Search we exert the statistic $V(\alpha, F_n)$, an adaptation of Butler et al (1993),

$$V(\alpha, F_n) = \frac{|k\hat{\mathbf{\Sigma}}_\alpha[F_n]|}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)} \int_0^{r_\gamma} r^{p+1}\phi'(r^2)dr)^{2p}} \,, \tag{4.1}$$

for

$$k = \begin{cases} 1 & \text{if } n < 30 \\ (1-\alpha) & \text{otherwise} \end{cases},$$

yielding an appropriate measure of the asymptotic variance for the location described by each subset retained after a proportion $\alpha = (1-\gamma)$ has been trimmed. $\hat{\mathbf{\Sigma}}_\alpha[F_n]$ is the corresponding sample covariance matrix, Γ is the usual gamma function, $\Gamma(v) = \int_0^\infty s^{v-1}e^{-s}ds$, $r_\gamma^2 = \chi^2_{1-\alpha,p}$ and $\phi(u) = (1/(2\pi))^{p/2}e^{-u/2}$. Recalling the Fisher Information argument, any reduction in information, the trimming of a data set for example, will be associated with a corresponding increase in the variance of parameter estimates, *if the data being removed is not extraneous.* This measure of asymptotic variance is expected to increase in inverse proportion to the size of the subsets it is applied to *unless* the subset retained, S_γ of size γn for some $0.5 \leq \gamma \leq 1$, is outlier free with respect to every other subset of data $S_{(x \geq \gamma)}$.

If the sample does not consist of outliers it has been observed, at least empirically, that there should be no minima of (4.1) occurring for an $\alpha > 0$. When the sample is contaminated by linear or radial outliers a solitary minimum is likely to occur when these outliers have been removed from the data set, indeed in 99% of simulations this minimum was a *global* minimum.

It so happens, see section 2.4.1, that when data sets are contaminated by outlying clusters, there can occur multiple minima. For bivariate data sets of size $n = 100$, when a proportion, $\epsilon = 0.3$, of the data is shifted about an outlier mean $d = 2\sqrt{\chi^2_{0.975,2}}$, there occurred multiple minima for an $\alpha > 0$ for between 15-20% of data sets. As the dimension increased these instances were less frequent, for example with regard to 4 dimensional data sets of size $n = 100$, with outlying proportion $\epsilon = 0.3$, multiple minima occurred on less than 10% of occasions. Equivalently, in over 90% of simulations there occurred *one* minimum for an $\alpha > 0$ in the vicinity of the correct cluster proportion. In section 2.4, where (4.1) underwent some preliminary tests, it was noticed that when 2 outlying clusters were present, $\epsilon_{C1} = \epsilon_{C2} = 0.2$, centred about *equal* distances either side of the main data, the response of the new proposal was again very accurate. The crucial aspect of these contaminations was the ensuing *minimum* of the multiple minima, occurring for

$\alpha > 0$, coincided with the greatest α for which a minimum occurred.

When multiple minima, m_i, $i = 1, \ldots, j$, of $V(\alpha, F_n)$ occur for an *increasing* $\alpha > 0$, we will have a series of j retained subsets $S_{\gamma_{\mathsf{m}_1}}, \ldots, S_{\gamma_{\mathsf{m}_j}}$, the smallest of which corresponding to $S_{\gamma_{\mathsf{m}_j}}$, when the particular observations identified as outlying for each proportion α are trimmed. This Chapter addresses those cases where

$$S_{\gamma_{\min_i(\mathsf{m}_i)}} \neq S_{\gamma_{\mathsf{m}_j}} \tag{4.2}$$

When (4.2) occurs the sample appears to satisfy one of two particular cases:

1. The sample is clustered.
2. The sample is clustered and has stray outliers.

Case 1 will compose a main cluster and at least *two* other clusters centred about different *absolute* displacements from the main mean. Case 2 describes the instances when stray observations with respect to a clustered configuration require trimming.

The phenomena of multiple minima of (4.1) occurring for $\alpha > 0$ was initially treated by choosing the outlying proportion α to be that α corresponding to a *minimum* of these $i = 1, \ldots, j$ minima, thus retaining the subset

$$S_{\gamma_{\min_i(\mathsf{m}_i)}}.$$

This selection strategy, for samples with one outlying cluster, coincided with choosing that minimum corresponding to the largest $\alpha > 0$ of data to be trimmed, equivalently subset

$$S_{\gamma_{\mathsf{m}_j}},$$

the *smallest* subset of data to be retained. This relationship between *minimum minima* for $\alpha > 0$ being equivalent to the greatest $\alpha > 0$ for which a minimum occurred was violated, thus satisfying (4.2), only when the data sets were of type Case 1 or Case 2. For example, see Figures 2.28-2.29 in section 2.5, where we chose that subset corresponding to $S_{\gamma_{\mathsf{m}_j}}$ as the final *trimmed* data set.

Section 2.6 provided very strong evidence to suggest that, in the event of multiple minima for $\alpha > 0$, *cleaning* the sample of those outliers forcing the minimum minima is not

sufficient if there are further minima, unaccounted for, corresponding to smaller subsets. Simulations have shown that the most effective strategy is to reapply until no minima occur. When this was carried out the samples appeared to have been been reduced to normally distributed, unimodal data sets and the true nature of the original data set revealed.

4.1 Example using an artificial data set

One helpful example data set is depicted in Figures 4.1-4.2, a $p = 3$ dimensional sample of size $n = 500$. This sample is composed of 5 clusters:

- Cluster 1: Sample proportion $\epsilon_{\text{C1}} = 0.45$, its 3 variables, $p_1,\ p_2,\ p_3$ distributed $N(0, 1)$.
- Cluster 2: $\epsilon_{\text{C2}} = 0.2$, $p_1, p_2 \sim N(0, 1),\ p_3 \sim N(5\sqrt{\chi^2_{0.975,3}}, 1)$.
- Cluster 3: $\epsilon_{\text{C3}} = 0.2$, $p_1, p_2 \sim N(0, 1),\ p_3 \sim N(-2.5\sqrt{\chi^2_{0.975,3}}, 1)$.
- Cluster 4: $\epsilon_{\text{C4}} = 0.1$, $p_1, p_2, p_3 \sim N(\sqrt{\chi^2_{0.975,3}}, 0.1)$, a *point mass* cluster.
- Cluster 5: $\epsilon_{\text{C4}} = 0.05$, $p_1 \sim N(0, 1),\ p_2 \sim N(4\sqrt{\chi^2_{0.975,3}}, 0.1),\ p_3 \sim N(1.5\sqrt{\chi^2_{0.975,3}}, 0.1)$, effectively a *line* mass cluster.

Figures 4.1-4.2 provide two perspectives on this data set, the second revealing a definitive cluster pattern. Figures 4.3-4.4 demonstrate the importance of *cleaning* a data set of those observations responsible for the *minimum* of any minima of the objective function occurring for $\alpha > 0$. Figure 4.3 exhibits minima occurring in the vicinity of sample proportions corresponding with cluster exclusions. Notice the minimum occurring at $n = 375$, this is due to the removal of clusters C3 and C5, when the Forward Search reaches $n = 425$ cluster C3 is restored to the subset of retained data whilst C4 has been removed along with C5. The algorithm for cluster detection simply involves retaining the subset of data, $S_{\gamma_{\min_i(\mathtt{m}_i)}}$, corresponding to the *minimum* of the minima and *reapplying* the **T2** proposal if further minima exist for smaller subsets. Figure 4.4 depicts the outcome when **T2** is reapplied after clearing the sample of clusters C4 and C5. We can see two minima

corresponding to the remaining clusters, C1, C2 and C3, indeed the precise minimum at $n = 375$ *disappears*, since C4 has been removed, with the second application of **T2**. As expected when **T2** was applied to the final retained subset of size $n = 225$, see Figure 4.4, there occurred no minima for $\alpha > 0$.

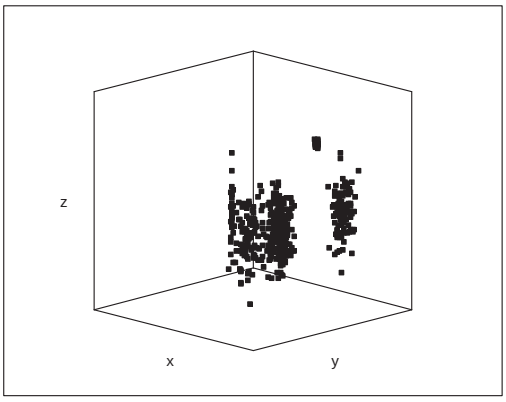


Figure 4.1: 3 dimensional perspective showing one outlying cluster.

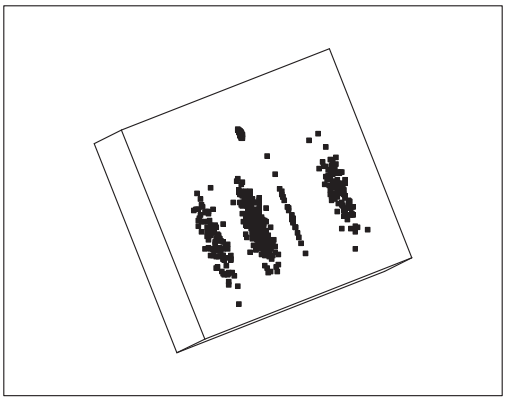


Figure 4.2: Perspective revealing exact cluster configuration.

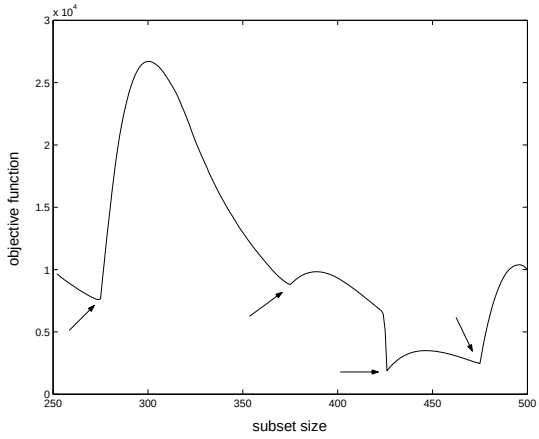


Figure 4.3: First application.

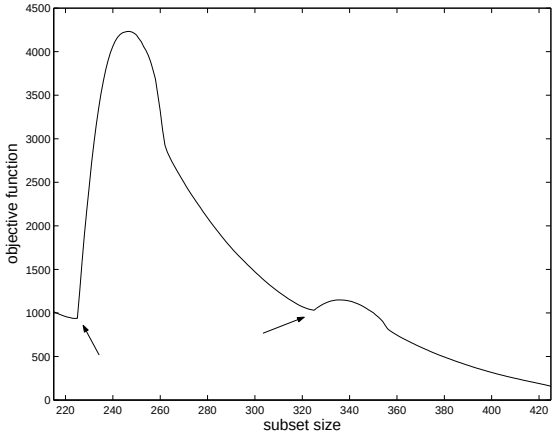


Figure 4.4: Second application after cleaning sample.

4.2 Simulations involving clustered data

This chapter introduces the notion of using the **T2** proposal to divulge the structure of a data set. **T2** can identify outliers in one application but an examination of the values of (4.1) for every subset selected by the Forward Search algorithm can potentially expose the exact configuration of clustering if clusters are present.

The first Monte Carlo session for this chapter will assess the performance of **T2** on two types of clustered data sets, $\mathbf{C622_{100,3}}$ and $\mathbf{C631_{100,3}}$ type samples. For example, Type $\mathbf{C631_{100,3}}$, see Table 4.1, refers to samples of size $n = 100$, dimension $p = 3$ with a proportion, $\epsilon_{C1} = 0.6$, of the data distributed $N_3(\mathbf{0}, I_3)$ and a proportion, $\epsilon_{C2} = 0.3$, of the data has its 3rd variable distributed $N(2d, 1)$ where $d = \sqrt{\chi^2_{0.975,3}} = 3.0575$. The final proportion, $\epsilon_{C3} = 0.1$ of the sample has its 1st variable distributed $N(0, 0.1)$ and its 2nd and 3rd variables distributed $N(\pm 4d, 0.1)$, respectively, which constitutes a *point mass* outlier.

Table 4.2 contains the proportion $\hat{P}_S$ of samples for which 3 clusters were identified along with the average cluster proportions detected, $\hat{P}_{C1}, \hat{P}_{C2}$ and $\hat{P}_{C3}$ which, ideally, should be:

- $P_{C1} = 0.6$, $P_{C2} = 0.2$ and $P_{C3} = 0.2$ for $\mathbf{C622_{100,3}}$.
- $P_{C1} = 0.6$, $P_{C2} = 0.3$ and $P_{C3} = 0.1$ for $\mathbf{C631_{100,3}}$.

For example, regarding the $\mathbf{C631_{100,3}}$ type samples, Table 4.2 shows 3 clusters were detected in 99% of simulations and the average sample proportion identified as belonging to each cluster in the proximity of the true planted proportions.

A pictorial example of sample types $\mathbf{C622_{100,3}}$ and $\mathbf{C631_{100,3}}$ is given in Figures 4.5-

Type	1st variable	2nd variable	3rd variable
$\mathbf{C622_{100,3}}$	$60N(0,1)$	$60N(0,1)$	$60N(0,1)$
	$20N(0,1)$	$20N(0,1)$	$20N(\mathbf{4d},1)$
	$20N(0,1)$	$20N(0,1)$	$20N(\mathbf{-2d},1)$
$\mathbf{C631_{100,3}}$	$60N(0,1)$	$60N(0,1)$	$60N(0,1)$
	$30N(0,1)$	$30N(0,1)$	$30N(\mathbf{2d},1)$
	$10N(0,\mathbf{0.1})$	$10N(\mathbf{-4d, 0.1})$	$10N(\mathbf{4d, 0.1})$

Table 4.1: Sample types $\mathbf{C622_{100,3}}$ and $\mathbf{C631_{100,3}}$.

4.6 and 4.9-4.10 whilst Figures 4.7-4.8 and 4.11-4.12 plot of the size of (4.1) with one and two applications, respectively. Figures 4.7-4.8 and 4.11-4.12 show the importance of thoroughly examining the output of **T2** for all subsets chosen by the Forward Search.

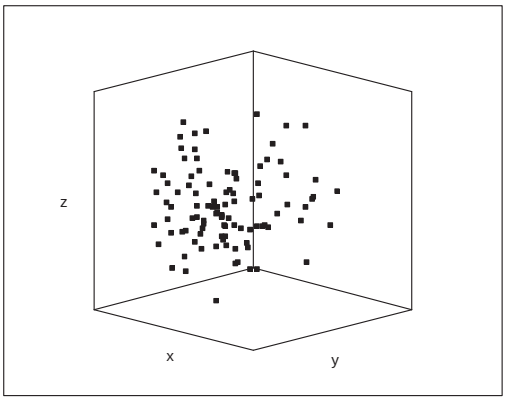


Figure 4.5: 3 dimensional **C622** perspective showing no obvious clustering.

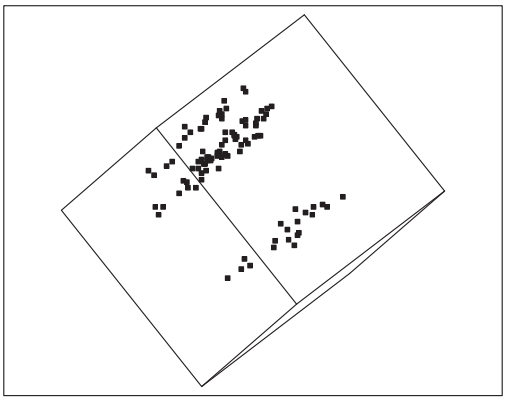


Figure 4.6: **C622** perspective revealing cluster configuration.

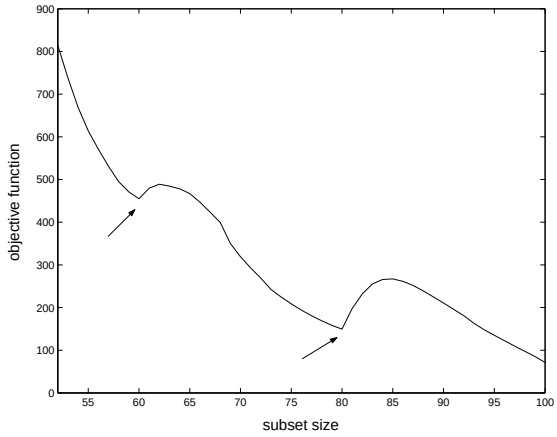


Figure 4.7: **C622** first application.

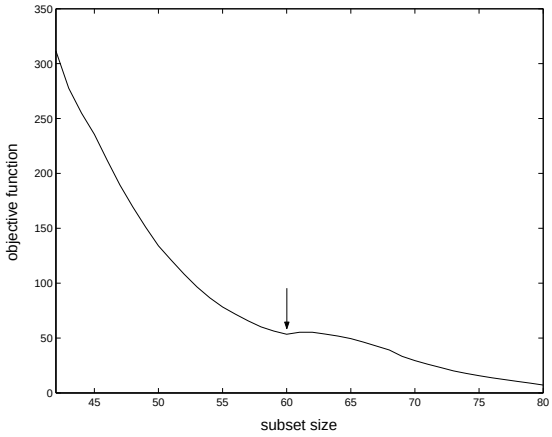


Figure 4.8: **C622** second application after cleaning sample.

The next series of Monte Carlo trials involved $p = 5$ dimensional data sets of size $n = 500$. The sample types investigated were $\mathbf{C532_{500,5}}$ and $\mathbf{C541_{500,5}}$, the latter containing a point mass cluster, see Table 4.3.

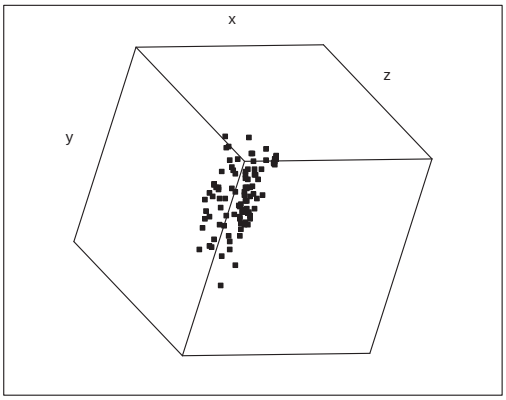


Figure 4.9: 3 dimensional **C631** perspective showing no obvious clustering.

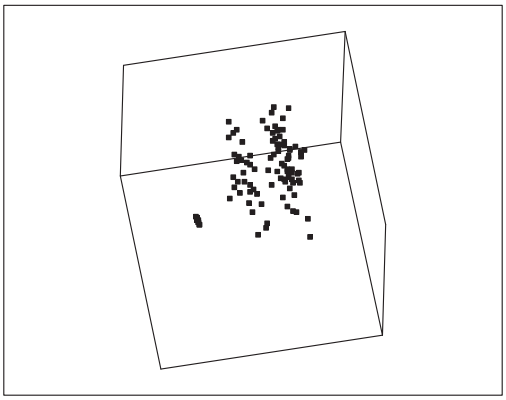


Figure 4.10: **C631** perspective revealing cluster configuration.

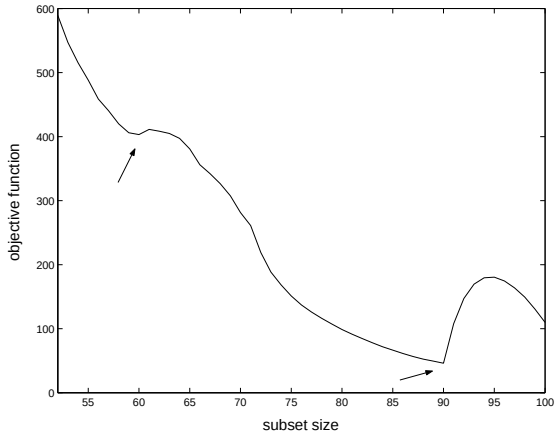


Figure 4.11: First application.

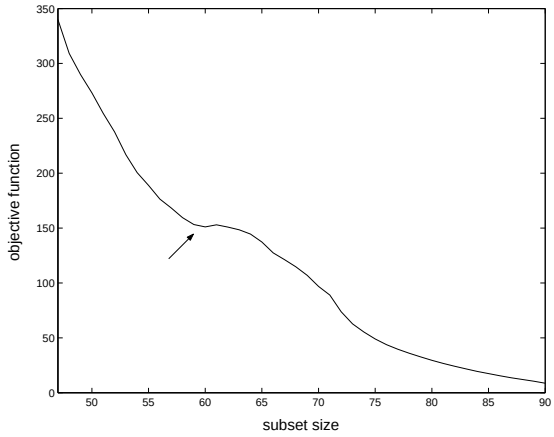


Figure 4.12: Second application after cleaning sample.

The simulations for $\mathbf{C532_{500,5}}$ and $\mathbf{C541_{500,5}}$ were conducted for a range of outlying mean displacements $d = 5, \ldots, 15$, noticing $\sqrt{\chi^2_{0.975,5}} = 3.5822$. Figures 4.13-4.14 depict the detection rates per sample, for example, cluster C2 in samples of type $\mathbf{C532_{500,5}}$ were detected in 31% of samples when the mean displacement of its corrupt variable was $d = 6$. When $d \geq 7$ the 3 clusters were nearly always detected. Table 4.4 contains the average sample proportions, over all $d = 5, \ldots, 15$, identified for each cluster when detected.

sample Type	$\hat{P}_S$	$\hat{P}_{\text{C1}}$	$\hat{P}_{\text{C2}}$	$\hat{P}_{\text{C3}}$
$\mathbf{C622_{100,3}}$	0.97	0.5965	0.2010	0.2025
$\mathbf{C631_{100,3}}$	0.99	0.6070	0.3002	0.1029

Table 4.2: Cluster detection proportions.

Sample Type	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$
$\mathbf{C532_{500,5}}$	$250N(0,1)$	$250N(0,1)$	$250N(0,1)$	$250N(0,1)$	$250N(0,1)$
	$150N(0,1)$	$150N(0,1)$	$150N(0,1)$	$150N(0,1)$	$150N(\boldsymbol{d},1)$
	$100N(\boldsymbol{d},1)$	$100N(0,1)$	$100N(0,1)$	$100N(0,1)$	$100N(0,1)$
$\mathbf{C541_{500,5}}$	$250N(0,1)$	$250N(0,1)$	$250N(0,1)$	$250N(0,1)$	$250N(0,1)$
	$200N(0,1)$	$200N(0,1)$	$200N(0,1)$	$200N(0,1)$	$200N(\boldsymbol{d},1)$
	$50N(-\boldsymbol{d},\mathbf{0.1})$	$50N(0,\mathbf{0.1})$	$50N(0,\mathbf{0.1})$	$50N(0,\mathbf{0.1})$	$50N(0,\mathbf{0.1})$

Table 4.3: Sample types $\mathbf{C532_{500,5}}$ and $\mathbf{C541_{500,5}}$.

sample Type	cluster	planted cluster proportion	average detected cluster proportion
$\mathbf{C532_{500,5}}$	C1	0.5	0.4980
	C2	0.3	0.3015
	C3	0.2	0.2007
$\mathbf{C541_{500,5}}$	C1	0.5	0.4976
	C2	0.4	0.4023
	C3	0.1	0.1003

Table 4.4: Cluster detection proportions.

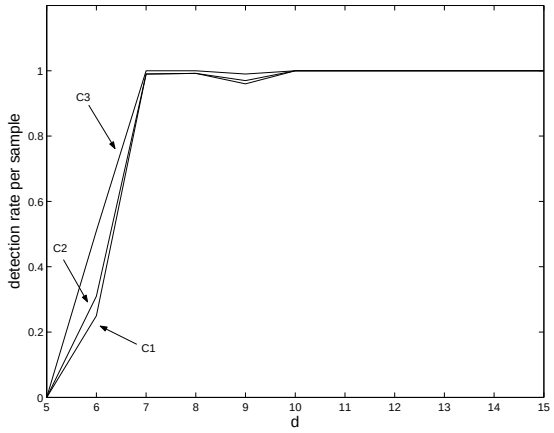


Figure 4.13: **C532** detection rates.

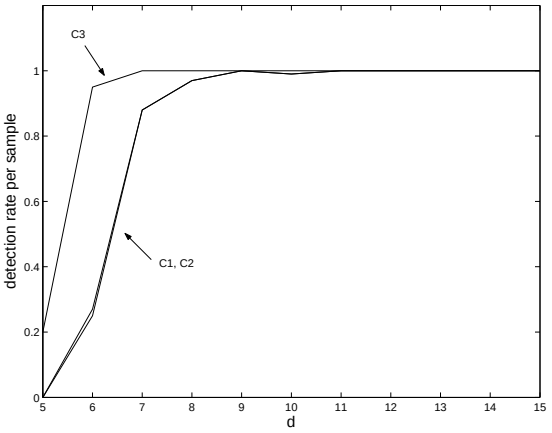


Figure 4.14: **C541** detection rates.

4.2.1 Relaxing breakdown restrictions

The first step of the **T2** proposal, as described in Chapter 2, is calculating a robust MCD estimate for location and scale using a variation on an algorithm devised by Woodruff and Rocke (1993).

The MCD estimate for location and scale is the corresponding mean, $\hat{\boldsymbol{\mu}}(S_{h/n})$ and covariance matrix, $\hat{\boldsymbol{\Sigma}}(S_{h/n})$, of the sample of size $h = \lfloor (n+p+1)/2 \rfloor$ resulting in the smallest covariance determinant with respect to *all* subsets of size h. This h was chosen to maximize the breakdown of the estimate. A sample breakdown point of an estimator T at S was defined earlier in section 1.4. It can also be interpreted, using the asymptotically equivalent expression, as

$$\epsilon^*(T) = \max\{(1-\xi) : \sup_{S_\xi} ||T(S_\xi)|| < \infty\}$$

where S_ξ is any subset obtained from S after replacing $n(1-\xi)$ points in S with arbitrary values. The *maximum* breakdown corresponds to the largest possible proportion of a sample for which there is a bound on an estimate when that proportion is corrupted without restriction. Of course this will fail to cope with a *bimodal* data set, **C55** say, composed of two clusters sharing equal proportion of the sample, $\epsilon_{C1} = \epsilon_{C2} = 0.5$, when

$p > 1$.

Sample types such as **C433** are also vulnerable to having their cluster configuration masked by the ensuing breakdown,

$$\epsilon^* = 1 - \frac{\lfloor (n+p+1)/2 \rfloor}{n}, \tag{4.3}$$

since **T2** involves measuring the asymptotic variance of subsets selected by a Forward Search beginning with the MCD subset of size $\lfloor (n+p+1)/2 \rfloor$. When applied to $\mathbf{C433_{100,3}}$ type samples it was discovered that on 5% of occasions the main cluster, of proportion $\epsilon_{\mathrm{C1}} = 0.4$, was detected with a first application of **T2**. In this case the second application of **T2** encounters a subset of size $0.6n$ composed of two clusters, C2 and C3, sharing equal proportion.

The breakdown is at *most* given by (4.3) since any subset of p points, in general position (for example no three points lie on the same line), must lie in a $(p-1)$ dimensional hyperplane H. If one was to corrupt $\lfloor (n-p+1)/2 \rfloor$ observations by placing them on this hyperplane we effectively have a subset of $\lfloor (n+p+1)/2 \rfloor = h$ observations with a determinant of zero. This corrupt subset of observations possesses the *smallest* possible determinant and will therefore comprise the MCD estimate.

The search algorithm we use to estimate the MCD begins with a subset, $S_{(p+1)/n}$, of $(p+1)$ points and is inflated point by point to contain h observations, $S_{\frac{\lfloor (n+p+1)/2 \rfloor}{n}}$. At each point of inflation the Mahalanobis distances, $M_i = \sqrt{(\mathbf{X}_i - \hat{\boldsymbol{\mu}}(S_\xi))^\top \hat{\boldsymbol{\Sigma}}\,(S_\xi)^{-1}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}(S_\xi))}$, of all $i = 1, \ldots, n$ observations with respect to subset $S_{\frac{(p+1)}{n} \leq \xi < \frac{\lfloor (n+p+1)/2 \rfloor}{n}}$ are calculated and those $(\xi n + 1)$ observations with the smallest M_i compose the inflated $S_{(\xi+1/n)}$.

The identification of clusters may not be an exercise towards estimating a *single* location and scale for a sample, rather we may be dealing with multi-modal data sets which may or may not have a cluster representing a *majority* subset of data. There is no penalty to the analysis if *all* observations belonging to a particular cluster, say subset $S_{\xi < \frac{\lfloor (n+p+1)/2 \rfloor}{n}}$, *do* lie on a hyperplane and do occupy the entire space, say $\lfloor (n+1)/2 \rfloor \approx 50\%$, covered by the covariance determinant.

It was decided to assess the performance of **T2**, when adhering to the MCD convention of seeking a minimum determinant for a subset of the size corresponding to maximum breakdown, when applied to sample types $\mathbf{C433_{100,3}}$ and $\mathbf{C55_{100,3}}$, see Table 4.5. The results were then compared with those obtained when choosing that subset of data, of size $\tilde{h} = \lfloor (n+1)/2 \rfloor$, minimizing the determinant as the starting point in the Forward Search. The Forward Search again searched for subsets, from this new starting point, that minimized (4.1).

The results of the Monte Carlo simulations used for this comparison are shown in Table 4.6. The results are not surprising, especially with regard to $\mathbf{C55_{100,3}}$ type samples, both clusters, C1 and C2, evading detection in all simulations using the conventional MCD subset as the starting point for the Forward Search. Using the smaller subset size, thus ignoring breakdown restrictions, the proportion of simulations, $\hat{P}_S$, for which the 2 clusters were identified was over 80%.

For the purposes of this thesis, ideally we apply **T2** using the conventional MCD estimate as a first step. Tables 4.5-4.6 show us that if we *suspect* clustering the starting point for the Forward Search can be any subset of size $\lfloor (n+1)/2 \rfloor$, $S_{\underline{\lfloor (n+1)/2 \rfloor}}$, minimizing the covariance determinant over all subsets of size $\lfloor (n+1)/2 \rfloor$.

Sample Type	$p=1$	$p=2$	$p=3$
$\mathbf{C433_{100,3}}$	$60N(0,1)$	$60N(0,1)$	$60N(0,1)$
	$30N(0,1)$	$30N(0,1)$	$30N(\mathbf{4}\boldsymbol{d},1)$
	$30N(0,1)$	$30N(-\mathbf{2}\boldsymbol{d},1)$	$30N(\mathbf{2}\boldsymbol{d},1)$
$\mathbf{C55_{100,3}}$	$50N(0,1)$	$50N(0,1)$	$50N(0,1)$
	$50N(0,1)$	$50N(0,1)$	$50N(-\mathbf{2.5}\boldsymbol{d},1)$

Table 4.5: Sample types $\mathbf{C433_{00,3}}$ and $\mathbf{C55_{100,3}}$.

For the depiction of two $\mathbf{C433_{100,3}}$ possibilities, when applying **T2**, observe Figures 4.15-4.20.

Figures 4.15-4.16 show us an example of such a sample type whilst Figures 4.17-4.18 show us a scenario which would be described using either subset types, discussed above, as Forward Search starting points. Figures 4.19-4.20 concern a scenario which occurred in

Initial subset size of Forward Search	$\hat{P}_S$	
	C433	**C55**
$\lfloor (n+p+1)/2 \rfloor$	0.96	< 0.01
$\lfloor (n+1)/2 \rfloor$	> 0.99	0.81

Table 4.6: Simulation results comparing different **T2** Forward Search starting points.

$\approx$ 5% of applications of **T2** to such data sets, the majority sample cluster is located first, see Figure 4.19. The two minor clusters can only be located, after clearing the data set of its main cluster, when the Forward Search begins with a subset, $S_{\underline{\lfloor (n+1)/2 \rfloor}}$, of size $\lfloor (n+1)/2 \rfloor$.

4.3 Example using real data

The 3-dimensional plots, Figures 4.21-4.22, concern measurements on 38 1978-79 model automobiles. The gas mileage in miles per gallon, MPG, as measured by Consumers' Union on a test track. The Horsepower, HP, and Displacement of the car (in cubic inches) as reported by automobile manufacturer (Reference: Henderson, H. V. and Velleman, P. F. (1981), "Building Regression Models Interactively." Biometrics, 37, 391-411. Data originally collected from Consumer Reports http://lib.stat.cmu.edu/DASL/Datafiles/Cars.html).

The original data set used here consists of an outlying cluster, denoted by the crosses. To give an example of why we need to reapply **T2** in the case of multiple minima to determine the structure of the data, one of the observations has been corrupted. The figures pertaining to the Chevette have been displaced, rendering it a stray point denoted by the plus sign, lying beyond the outlying cluster.

Figures 4.23-4.24 divulge the size of the objective function when **T2** is applied, first to the whole data set, then after the corrupted point is removed.

This example can lead to the conjecture that the corrupted observation, detected with

one application of **T2**, was the *sole* outlier with respect to a *clustered* sample, and once removed, the clustered configuration was confirmed with a second application of **T2**.

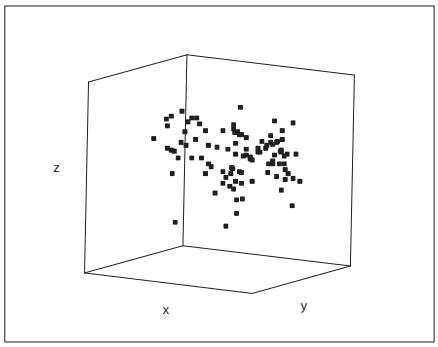


Figure 4.15: **C433** perspective showing no obvious clustering.

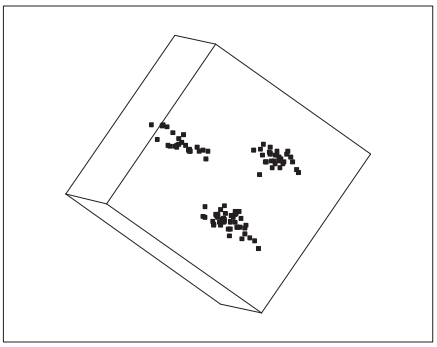


Figure 4.16: Perspective showing clusters.

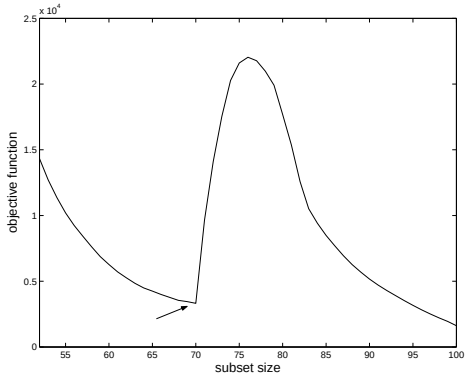


Figure 4.17: **C433** First application of **T2** detects a minor cluster.

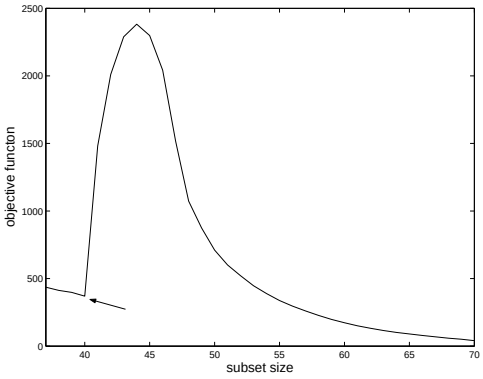


Figure 4.18: Second application of **T2** revealing other two clusters.

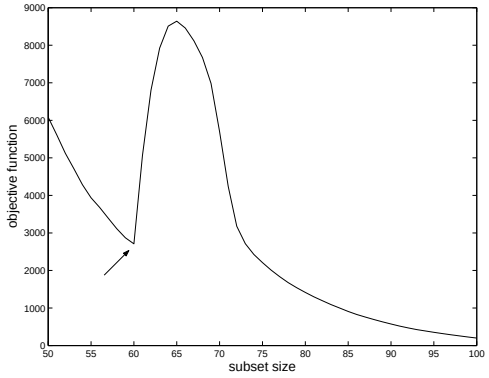


Figure 4.19: **C433** First application of **T2** isolates main cluster.

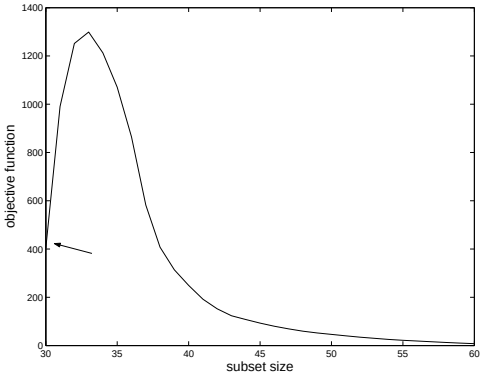


Figure 4.20: Second application after loosening breakdown restrictions.

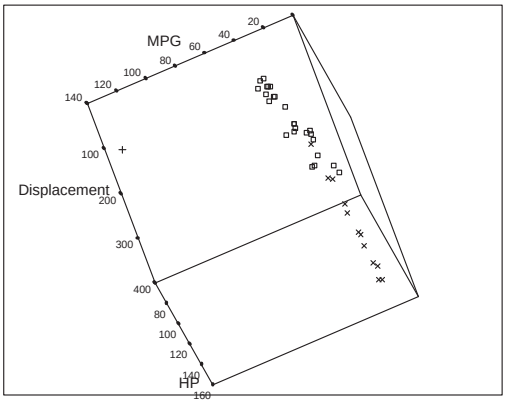


Figure 4.21: Cars perspective exposing planted outlier.

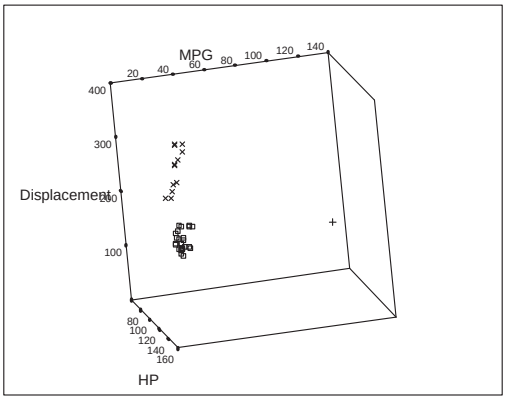


Figure 4.22: Cars perspective exposing outlying cluster.

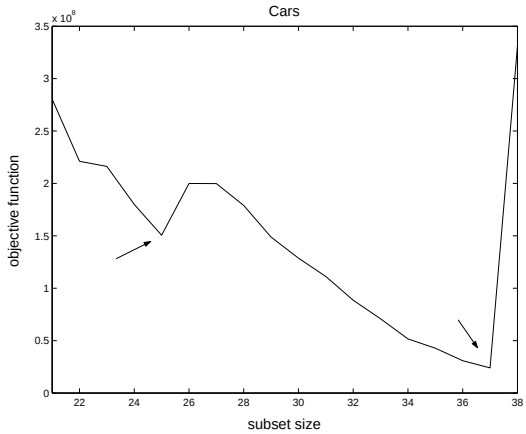


Figure 4.23: Multiple Minima

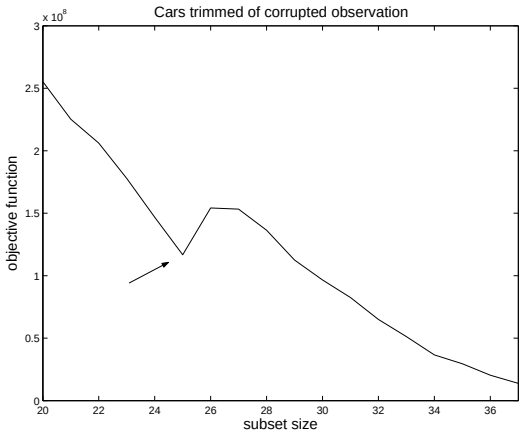


Figure 4.24: Stray point removed.

Chapter 5

Other Diagnostics

5.1 Principal Components Analysis

The analysis of multivariate data is an attempt to find patterns or relationships in the data, the observations identified as outlying are not consistent with these patterns or relationships. Principal Components Analysis, (PCA), is a powerful tool in highlighting patterns in high dimensional data and can be used to compress data with a minimal loss of information. With a PCA we transform an original set of correlated variables into a set of uncorrelated variables. This new set of uncorrelated variables are called *Principal Components*, and are composed of orthogonal linear combinations of the original variables, the principal components form an orthogonal basis for the data space. (Chatfield and Collins 1980, Johnson and Wichern 1998). Each of these principal components contribute to the overall variability of the data, the objective of a PCA is to hopefully find a number, $l < p$, of Principal Components that explain most of the variability. If fewer than p principal components account for most of the variability, the effective dimensionality of the problem is less than p and the data set can be simplified with minimal loss of information.

Given the p-dimensional random vector $\mathbf{X}^\top = (X_1, X_2, \ldots, X_p)$ we consider the linear

combinations (Johnson and Wichern 1998),

$$\begin{aligned} Y_1 &= \mathbf{a_1}^\top \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= \mathbf{a_2}^\top \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{a_p}^\top \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned}$$

We want these principal components $Y_1, Y_2, \ldots, Y_p$ to be uncorrelated, $\mathrm{Cov}(Y_i, Y_k) = 0$, which will be the case if the $\mathbf{a}_i$ are orthogonal, in fact the transformation will be an orthogonal rotation in p-space when this is the case (Chatfield and Collins 1980). It can be seen that

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\mathbf{a}_i^\top \mathbf{X}) = \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{a}_i \;\; i = 1, 2, \ldots, p,$$

and

$$\mathrm{Cov}(Y_i, Y_k) = \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{a}_k \;\; i \neq k.$$

The first principal component is that linear combination which has the maximum variance so we add the constraint that $\mathbf{a}_1^\top \mathbf{a}_1 = 1$ otherwise the $\mathrm{Var}(Y_i) = \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{a}_i$ can be arbitrarily large. Similarly we have constraints $\mathbf{a}_i^\top \mathbf{a}_i = 1$ for all $i = 1, \ldots, p$.

If $\mathbf{B}$ is a $(p \times p)$ positive definite matrix with eigenvalues $\Lambda = \{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0\}$ corresponding to eigenvectors $\mathbf{V} = \{\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_p}\}$, matrix algebra results give us

$$\frac{\mathbf{x}^\top \mathbf{B} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \frac{\mathbf{x}^\top \mathbf{V} \mathbf{\Lambda^{1/2}} \mathbf{V}^\top \mathbf{V} \mathbf{\Lambda^{1/2}} \mathbf{V}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{V} \mathbf{V}^\top \mathbf{x}} = \frac{\mathbf{z}^\top \mathbf{\Lambda} \mathbf{z}}{\mathbf{z}^\top \mathbf{z}}$$

where $\mathbf{z} = \mathbf{V}^\top \mathbf{x}$. We can see from this that

$$\frac{\sum_{i=1}^p \lambda_i z_i^2}{\sum_{j=1}^p z_j^2} \leq \lambda_1 \frac{\sum_{i=1}^p z_i^2}{\sum_{j=1}^p z_j^2} = \lambda_1$$

since $\lambda_1 \geq \lambda_j \quad j = 2, \ldots, p$ and so

$$\max \frac{\mathbf{x}^\top \mathbf{B} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_1$$

when $\mathbf{x} = \mathbf{e_1}$ since $\mathbf{z} = \mathbf{V}^\top \mathbf{e_1} = (1 \;\; 0 \;\; \ldots \;\; 0)^\top$. Similarly

$$\max_{\mathbf{x} \perp \mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_k}} \frac{\mathbf{x}^\top \mathbf{B} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_{k+1} \;\text{ for }\; \mathbf{x} = \mathbf{e_{k+1}}, \;\; k = 1, 2, \ldots, p-1. \tag{5.1}$$

Further results show that if $Y_i = \mathbf{e}_i^\top \mathbf{X},\ i = 1, 2, \ldots, p$, are the principal components of a sample with covariance matrix $\mathbf{\Sigma}$ then

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i)$$

and the proportion of total variance explained by the kth principal component is therefore

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad k = 1, 2, \ldots, p. \tag{5.2}$$

If all the eigenvalues are distinct then the corresponding eigenvectors are orthogonal, or may be chosen to be so otherwise. We can therefore derive the principal components from the covariance matrix of the data set if we replace matrix $\mathbf{B}$ with $\mathbf{\Sigma}$, the ith Principal Component is therefore given by

$$Y_i = \mathbf{e_i}^\top \mathbf{X} \quad i = 1, 2, \ldots, p,$$

where $\mathbf{e}_i$ is the eigenvector decomposition of $\mathbf{\Sigma}$ so that Y_i becomes the mapping of the p dimensional data space, $\mathbf{X}$, onto the space defined by the p orthogonal eigenvectors. The original data is rotated such that the eigenvectors become the new axes, the eigenvector corresponding the the largest eigenvalue may be an approximation of the line of best fit. The other eigenvectors describe how the data deviates about this line and those eigenvectors that explain only a small amount of the variance can be ignored, thus reducing the dimensionality of the data set.

Although normality does not need to be assumed for a PCA, it can reveal important information about the ellipsoidal distribution. Indeed we can think of a PCA as an orthogonal rotation of the axis such that they line up with the the ellipsoids axis, the 1st principal component describes the major axis of the ellipsoid and approximates a regression line.

Once the eigenvalue-eigenvector pairs have been established from the p dimensional data set, we can describe the principal components and their corresponding contribution to the overall variance. If $k < p$ principal components explain, for example $> 80\%$ of the variance, we can replace the original p variables with the k principal components, effectively

projecting the original data set onto the reduced space spanned by the k eigenvectors, simplifying the data set. This projection can be assessed for clusters and stray outliers.

5.1.1 New PCA proposal and simulations

The new proposal PCA for this thesis will be assessed using Monte Carlo trials and compared with the classic PCA and the ROBPCA, where the latter is a robust PCA devised by Hubert, Rousseeuw and Branden (2003).

The simulations given here detail the performance of these three PCA methods when applied to normally distributed samples of size $n = 20, 50$ and 100. Each sample incurring three pre-specified levels of contamination, $\epsilon = 0, 0.1$ and 0.2 respectively. For each sample size, the clean data is generated with dimension $p = 4$ about $\boldsymbol{\mu}_1 = (0, 0, 0, 0)$ with

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 8 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.3}$$

(see Hubert, Rousseeuw and Branden 2003).

We tabulated the results for 4 cases of contamination for each proportion of outlying data respectively. These cases composed 4 different levels of displacement of the $4th$ variable whence the proportion of outlying data is generated about $\boldsymbol{\mu}_2 = (0, 0, 0, u)$, where $u = 6, 10, 15, 20$ respectively.

Due to computational expense, only one other type of data set was generated for this Monte Carlo analysis, data sets of size $n = 100$ and of dimension $p = 10$. For these samples the clean data was generated about the zero vector $\boldsymbol{\mu}_3$ of length 10 with

$$\boldsymbol{\Sigma}_2 = \text{diag}(21, 18, 15, 6/7, \ldots, 6/7).$$

The contaminated proportion was generated about the mean vector $\boldsymbol{\mu}_4 = (0, 0, 0, u, 0, 0, 0, 0, 0, 0)$ where $u = 6, 10, 15, 20$ for the 4 cases of corrupted data assessed.

By (5.2) when dealing with the data sets generated in dimension 4 according to $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_1)$ we expect,

$$\frac{\sum_{i=1}^{3} \lambda_i}{\sum_{i=1}^{4} \lambda_i} = 0.9333, \tag{5.4}$$

whilst for cases where 10 dimensional data sets are generated $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_2)$ we expect,

$$\frac{\sum_{i=1}^{3} \lambda_i}{\sum_{i=1}^{10} \lambda_i} = 0.9000, \tag{5.5}$$

and so by default we select the eigenvalues corresponding to the 3 largest eigenvalues for the principal components for the $p = 4$ and $p = 10$ dimensional data sets trialed. We can use the simulations to compare the relevant proportions of variability explained by these 3 PCA algorithms with the ideal proportions given in (5.4), (5.5).

The new PCA method T1PCA/T2PCA, *initially*, will simply involve the straight forward application of the new proposal, **T1** or **T2** depending on sample size, as the first step. This will *clean* the sample of any possible outliers before any PCA is undertaken. Here we establish from the outset an outlier free subset of the data to which we can apply a classic version PCA.

T1PCA:

Step 1: Apply new proposal **T1** to sample of size $n \leq 20$, removing any outliers detected.

Step 2: Conduct classical PCA.

T2PCA:

Step 1: Apply new proposal **T2** to sample of size $n > 20$, removing any outliers detected.

Step 2: Conduct classical PCA.

A preliminary comparison between the new T2PCA proposal and the classic approach to a PCA is depicted in Figures 5.1-5.3. The comparison was conducted for normally distributed samples of size $n = 100$, dimension $p = 4$ corrupted by a pre-specified proportion of the pth variable. Figures 5.1-5.3 portray, for an increasing outlier mean of the contaminated proportion, the amount of variability explained by the first 3 principal components. The dashed line represents the ideal proportion of variability the first 3 principal components should explain given perfectly distributed samples $N(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$.

The T2PCA has performed slightly better than the non-robust Classical PCA method. To

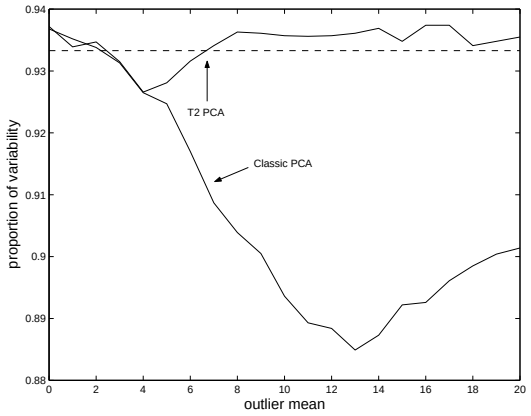


Figure 5.1: Proportion of variability, $n = 100,\ p = 4,\ \epsilon = 1/n$.

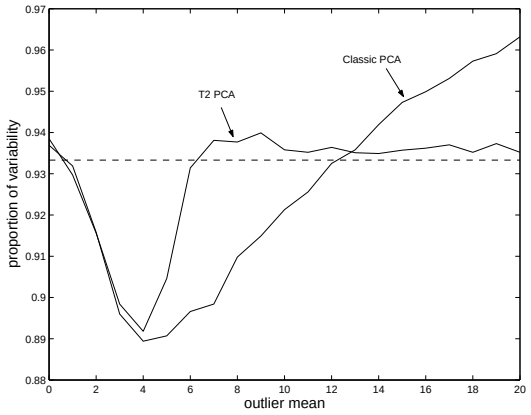


Figure 5.2: Proportion of variability, $n = 100,\ p = 4,\ \epsilon = 0.1$.

expose the strength of T2PCA against the Classical PCA we need an additional method of assessment.

When applying ROBPCA for data sets with $p < n$, the MCD estimate with $h = \lfloor 0.75n \rfloor$, hence $\epsilon^* = 1 - \frac{\lfloor 0.75n \rfloor}{n}$, can be obtained at the outset. From this set of $\lfloor 0.75n \rfloor$ points composing that set yielding the MCD, the mean $\hat{\boldsymbol{\mu}}_1$ and Covariance matrix $\hat{\boldsymbol{\Sigma}}_1$ is calculated. A preliminary PCA is performed using $\hat{\boldsymbol{\Sigma}}_1$ and the whole data set projected onto the subspace described by the first, pre-specified $k < p$ eigenvectors $\tilde{\boldsymbol{P}}_{p,k}$ (Hubert et al 2003),

$$\mathbf{X}_{n,k} = (\mathbf{X}_{n,p} - \mathbf{1}_n \hat{\boldsymbol{\mu}}_1^\top)\tilde{\boldsymbol{P}}_{p,k}.$$

A second MCD, $\gamma = \lfloor 0.75 \rfloor$ by default, is obtained from this projected data set from which

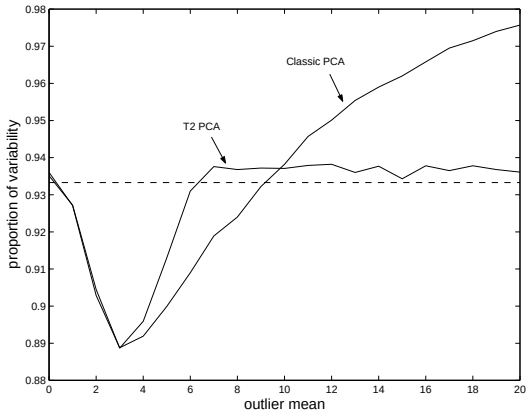


Figure 5.3: Proportion of variability $n = 100,\ p = 4,\ \epsilon = 0.2$.

is calculated $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\Sigma}}_2$. Using the correction factor first proposed by Rocke and Woodruff (1996),

$$c = \frac{\{d_{\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2}\}_{0.75n}}{\chi^2_{0.75,k}},$$

where d_i are the Mahalanobis distances with respect to $\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2$, the Mahalanobis distances, M_i, are then calculated for each projected observation with respect to $\hat{\boldsymbol{\mu}}_2$ and $c\hat{\boldsymbol{\Sigma}}_2$. The set of $M_i > \sqrt{\chi^2_{0.975,k}}$ are identified as outliers and scrapped from the analysis. The resulting covariance matrix, $\hat{\boldsymbol{\Sigma}}_3$, is then used to calculate new eigenvalue-eigenvector pairs $\boldsymbol{L}_{k,k}, \tilde{\boldsymbol{P}}_{k,k}$. The eigenvectors and the resulting mean, $\hat{\boldsymbol{\mu}}_3$, are transformed back to $\Re^p$ for the final mean, principal components and covariance matrix.

This transformation of the mean $\hat{\boldsymbol{\mu}}_3$ back to $\Re^p$ is calculated,

$$\hat{\boldsymbol{\mu}}_4 = \hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_3 \tilde{\boldsymbol{P}}_{p,k},$$

while the transformation of the principal components $\tilde{\boldsymbol{P}}_{k,k}$ back to $\Re^p$ is calculated via

$$\boldsymbol{P}_{p,k} = \tilde{\boldsymbol{P}}_{p,k} \tilde{\boldsymbol{P}}_{k,k}$$

(Hubert et al 2002).

The **T2** proposal was also used in conjunction with this ROBPCA. First we can carry out a ROBPCA, we then establish that subset of data minimizing out objective function once we have calculated the *second* MCD. When this was added as another step to the

ROBPCA it increased computational time and the results were not as impressive as when using the original ROBPCA algorithm. The *initial* T2PCA proposal discussed above was also superior to this more complicated methodology and also reduces the computational time needed for the full ROBPCA found in Hubert et al (2003).

Tables 5.1-5.3 contain the proportion of variabilities explained by the first 3 Principal Components using the three methods discussed above. The T2PCA proposal, introduced here, has again produced excellent results. Notice, for Tables 5.1-5.3, these assessed variability proportions do sometimes reach "impossible" levels, whenever the variability proportion covered by the first 3 PC's was calculated to be > 1.0. This is because the ROBPCA only estimates the first, pre-specified $k < p$ eigenvalues, k=3 in these scenarios, expected to be responsible for $\geq 80\%$ of variability. We have used the *ideal* total variability as the denominator which, by (5.3) is $\lambda_1 + \cdots + \lambda_4 = 15$, in the ensuing ratio with the 3 largest, derived, eigenvalues.

n	p	ϵ	u	Classic PCA	ROBPCA	T1PCA
20	4	0		0.9466	0.6797	0.9685
		0.1	6	1.1266	0.7794	1.0503
			10	1.5232	0.8018	0.9504
			15	2.3227	0.8004	0.9723
			20	3.4218	0.8834	0.9192
		0.2	6	1.2971	0.8797	1.1455
			10	2.019	0.9614	0.9583
			15	3.4264	0.9801	0.9158
			20	5.4019	1.1586	0.9457

Table 5.1: Expected proportion of variability explained 0.9333.

n	p	ϵ	u	Classic PCA	ROBPCA	T2PCA
50	4	0		0.934	0.7992	0.9349
		0.1	6	1.1088	0.8185	0.9625
			10	1.4909	0.8744	0.9392
			15	2.2532	0.8407	0.9464
			20	3.318	0.8755	0.9365
		0.2	6	1.2726	0.8343	0.9714
			10	1.968	0.8647	0.9058
			15	3.3216	0.9708	0.9413
			20	5.235	1.0152	0.9386

Table 5.2: Expected proportion of variability explained 0.9333.

Krzanowski's minimum angle

A further test which can assess the ability of these methods to protect a PCA from outlying data, is to calculate the angle between the resulting space spanned by the principal components and the space spanned by the ideal situation. An ideal sample is that perfectly generated, 4 dimensional data set, such that its eigenvalues would be equal to diag$(8, 4, 2, 1)$ and its eigenvectors $\boldsymbol{I}_4$, the Identity matrix composed of the 4 *ideal* direction cosines. Theorem 1 in Krzanowski (1979) states that the *minimum* angle between an arbitrary vector in k-space of principal components of one data set and the most nearly parallel vector to it in the k-space of the principal components of another data set, is

$$\arccos(\lambda^{1/2})$$

where λ is the largest eigenvalue of, for our purposes, $\boldsymbol{I}_{3,4}\boldsymbol{P}_{4,3}\boldsymbol{P}_{3,4}\boldsymbol{I}_{4,3}$ where $\boldsymbol{P}$ represents the eigenvectors for each sample assessed (Hubert et al 2003). Here we prefer to determine the *maximum* angle between the resulting space, described by the principal components, and the space defined by the ideal principal components $E_k = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$. This corresponds to calculating

$$\arccos(\lambda^{1/2})/(\pi/2)$$

where λ is the smallest eigenvalue of $\boldsymbol{I}_{3,4}\boldsymbol{P}_{4,3}\boldsymbol{P}_{3,4}\boldsymbol{I}_{4,3}$ and the division by $\pi/2$ standardizes this value (Hubert et al 2003).

The maximum angle between the ideal projection and the derived PCA are smaller for the larger displacements, $u = 10, 15, 20$, of the corrupt data when using the fast robust T1PCA/T2PCA method. This simpler method using the **T1** and **T2** proposals is weaker when applied to data sets, of dimension $p = 4$, suffering a contamination due to the smaller displacement of ϵ-proportion of data, i.e. $u = 6$. Tables 5.4-5.6 tabulate the average size of this maximum angle when using the 3 PCA methods being assessed here.

Figures 5.4-5.6 show the comparisons for maximum angle between the ideal projection and the derived projection using the 3 PCA methods discussed for samples of size $n = 100$ and dimension $p = 4$. These Figures show ROBPCA performing strongly with respect to

n	p	ϵ	u	Classic PCA	ROBPCA	T2PCA
100	4	0		0.9377	0.8108	0.9359
		0.1	6	1.094	0.8717	0.9601
			10	1.4769	0.8828	0.9310
			15	2.2373	0.862	0.9194
			20	3.2887	0.8872	0.9348
		0.2	6	1.2588	0.8783	0.9547
			10	1.9421	0.8907	0.9245
			15	3.297	0.9131	0.9174
			20	5.1794	0.9389	0.9146

Table 5.3: Expected proportion of variability explained 0.9333.

n	p	ϵ	u	Classic PCA	ROBPCA	T1PCA
20	4	0		0.2687	0.4329	0.2617
		0.1	6	0.8493	0.4199	0.5748
			10	0.9306	0.4176	0.2923
			15	0.956	0.3977	0.2982
			20	0.9699	0.4354	0.2915
		0.2	6	0.9076	0.4016	0.6205
			10	0.9506	0.3878	0.3295
			15	0.9688	0.3602	0.3792
			20	0.9772	0.3731	0.2982

Table 5.4: Average maximum angle

n	p	ϵ	u	Classic PCA	ROBPCA	T2PCA
50	4	0		0.1446	0.2785	0.1384
		0.1	6	0.9003	0.2618	0.3425
			10	0.9558	0.2833	0.1702
			15	0.974	0.2407	0.1550
			20	0.9813	0.2462	0.1740
		0.2	6	0.9438	0.2141	0.3448
			10	0.9696	0.2045	0.1788
			15	0.9817	0.1905	0.1609
			20	0.9867	0.2102	0.1897

Table 5.5: Average maximum angle

all the mean outlier displacements and the T2PCA spiking to poor values for the mean cluster displacements in the range $d = 4$ and $d = 5$. The poor behaviour at this level of displaced contamination corresponds with the results shown in Table 2.3, where a slight degree of undertrimming was advised by **T2** for clusters resulting from smaller shifts of outlier mean. T2PCA settles down for mean outlier shifts $d \geq 6$. The Classical PCA converges to the worst possible angle discrepancy between the ideal projection and that derived using this non-robust method.

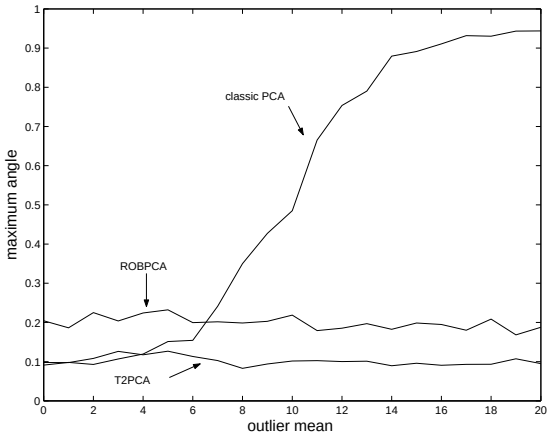


Figure 5.4: Maximum angle $n = 100,\ p = 4,\ \epsilon = 0.01$.

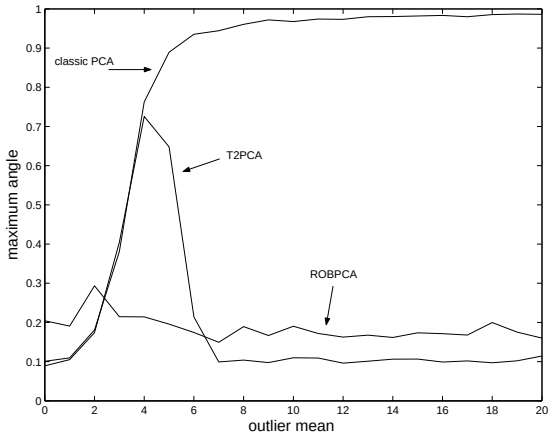


Figure 5.5: Maximum angle $n = 100,\ p = 4,\ \epsilon = 0.1$.

Tables 5.7-5.8 contain the results when the 3 PCA methods were applied to Monte Carlo

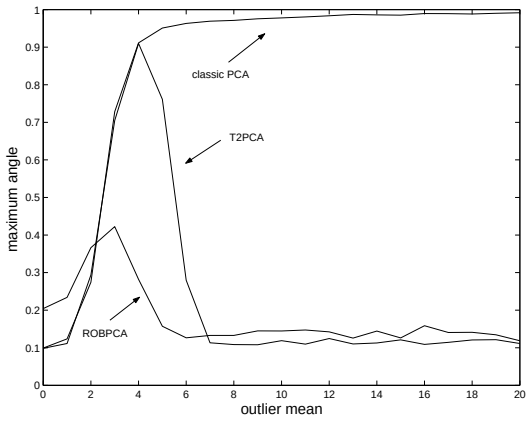


Figure 5.6: Maximum angle $n = 100,\ p = 4,\ \epsilon = 0.2$.

samples of size $n = 100$ and dimension $p = 10$. The T2PCA performs very strongly for all contamination types. One hopes the proportion of variability accounted for by the first 3 PC's is 0.9000 and the maximum angle as small as possible.

5.1.2 t_5-distributed data sets

The above analysis was also applied to samples distributed according to the multivariate t-distribution with 5 degrees of freedom incurring the same proportions of planted outliers. Figures 5.7-5.24 depict the comparisons between the 3 PCA methods. These Figures delineate the proportion of variability using the *exact* values calculated from the Classica1 PCA and the PCA proposal against the estimate derived, as above, for the ROBPCA. These Figures show the all round strength of the **T1** and **T2** PCA's. With regard to the proportion of variability explained by the first 3 PC's we see the estimate for ROBPCA behaving somewhat erratic and under representative of the ideal proportion expected. On the other hand the maximum angle between the new proposal's derived PCA and the ideal PCA is weaker than the corresponding ROBPCA values for lower levels of mean outlier displacement. For outlier means in excess of $d = 6$ the new PCA method converges to the ROBPCA values whilst the Classical PCA converges to the worst possible case. All these comparisons are consistent with the same comparisons conducted for normally distributed data sets.

Table 5.9 contains the results of the 3 PCA algorithms when applied to t_5-distributed samples of size $n = 100$, dimension $p = 10$ contaminated as per Tables 5.7-5.8. The T2PCA proposal is again very accurate regarding the proportion of variability is expected to be 0.9000. The average maximum angle between the derived projection and the ideal PCA projection is roughly the same as for the ROBPCA.

n	p	ϵ	u	Classic PCA	ROBPCA	T2PCA
100	4	0		0.0963	0.2043	0.1112
		0.1	6	0.9318	0.1744	0.2446
			10	0.9712	0.1903	0.1121
			15	0.983	0.1736	0.1000
			20	0.9871	0.1601	0.1017
		0.2	6	0.9593	0.1265	0.2031
			10	0.9793	0.1446	0.1088
			15	0.9871	0.1263	0.1035
			20	0.9904	0.1186	0.1226

Table 5.6: Average maximum angle

n	p	ϵ	u	Classic PCA	ROBPCA	T2PCA
100	10	0		0.9139	0.7969	0.9206
		0.1	6	0.9103	0.8066	0.8966
			10	0.9013	0.808	0.8971
			15	0.9273	0.7902	0.9075
			20	1.0667	0.8048	0.8900
		0.2	6	0.919	0.8119	0.9090
			10	0.9135	0.8164	0.8928
			15	1.0585	0.8051	0.9043
			20	1.3489	0.7985	0.9121

Table 5.7: Expected proportion of variability explained 0.9000.

n	p	ϵ	u	Classic PCA	ROBPCA	T2PCA
100	10	0		0.0537	0.079	0.0531
		0.1	6	0.0684	0.0724	0.0667
			10	0.1096	0.0749	0.0571
			15	0.4512	0.0802	0.0581
			20	0.8883	0.0742	0.0574
		0.2	6	0.0793	0.1091	0.0812
			10	0.2438	0.0692	0.0718
			15	0.8836	0.0677	0.0629
			20	0.947	0.0681	0.0614

Table 5.8: Average maximum angle.

n	p	ϵ	u	*Variability proportion*			*Maximum angle*		
				Classic PCA	ROBPCA	T2PCA	Classic PCA	ROBPCA	T2PCA
100	10	0		0.9143	0.6267	0.8812	0.0567	0.0795	0.0560
		0.1	6	0.8964	0.6486	0.8781	0.0695	0.0757	0.0684
			10	0.9096	0.6425	0.8431	0.12	0.0738	0.0604
			15	0.9341	0.6647	0.8473	0.5066	0.0699	0.0610
			20	1.1011	0.6762	0.838	0.9	0.0719	0.0590
		0.2	6	0.9064	0.6833	0.8928	0.0881	0.1237	0.0834
			10	0.9209	0.6706	0.8564	0.2781	0.1753	0.1392
			15	1.0823	0.6416	0.8539	0.8937	0.068	0.1059
			20	1.3568	0.6545	0.8695	0.9467	0.0687	0.1159

Table 5.9: Results of simulations for t_5 data sets of size $n = 100$ and dimension $p = 10$.

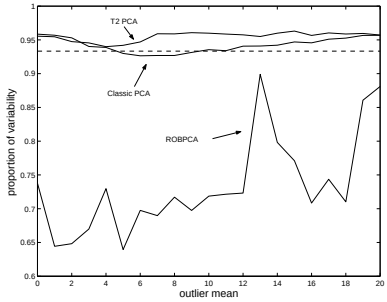


Figure 5.7: Proportion of variability explained $n = 20$ 1 outlier.

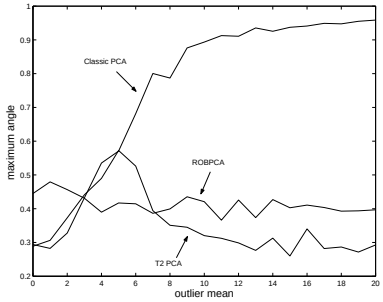


Figure 5.8: Maximum angle $n = 20$ 1 outlier.

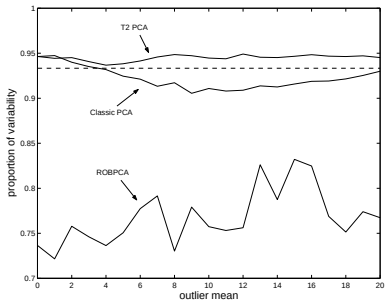


Figure 5.9: Proportion of variability explained $n = 50$ 1 outlier.

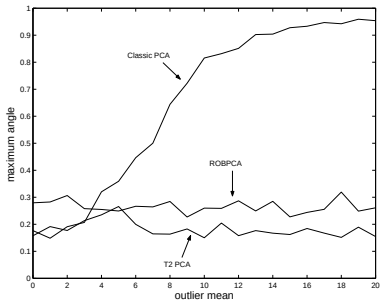


Figure 5.10: Maximum angle $n = 50$ 1 outlier.

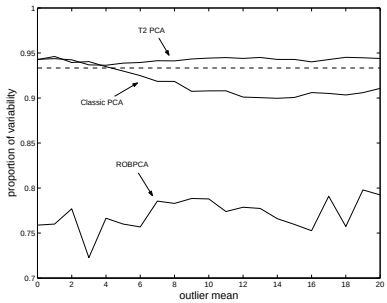


Figure 5.11: Proportion of variability explained $n = 100$ 1 outlier.

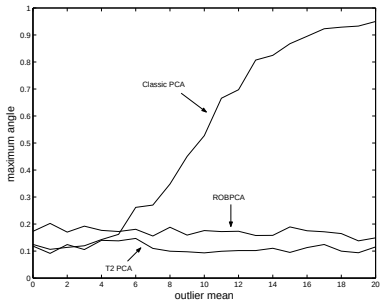


Figure 5.12: Maximum angle $n = 100$ 1 outlier.

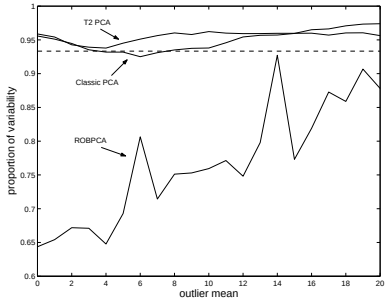


Figure 5.13: Proportion of variability explained $n = 20$ 2 outliers.

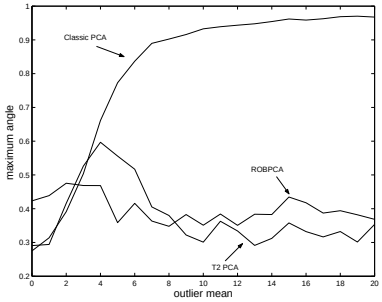


Figure 5.14: Maximum angle $n = 20$ 2 outliers.

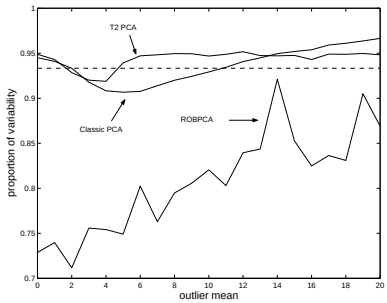


Figure 5.15: Proportion of variability explained $n = 50$ 5 outliers.

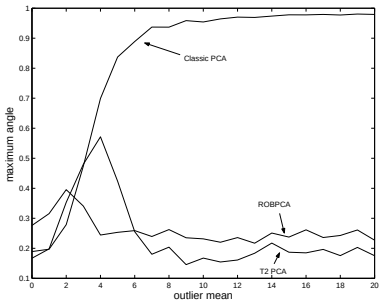


Figure 5.16: Maximum angle $n = 50$ 5 outliers.

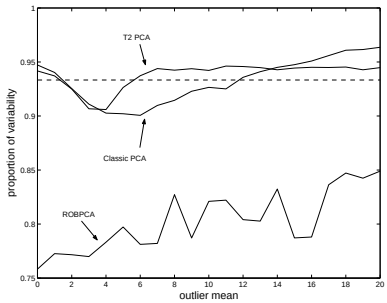


Figure 5.17: Proportion of variability explained $n = 100$ 10 outliers.

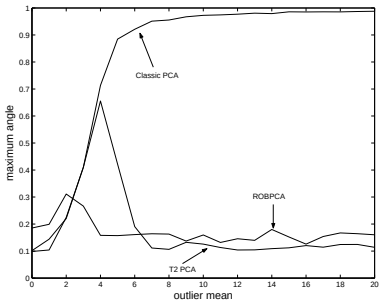


Figure 5.18: Maximum angle $n = 100$ 10 outliers.

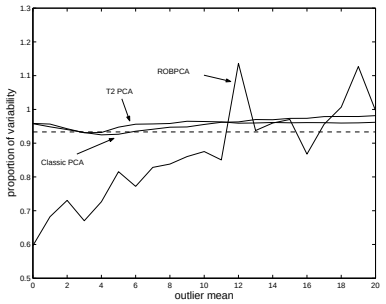


Figure 5.19: Proportion of variability explained $n = 20$ 4 outliers.

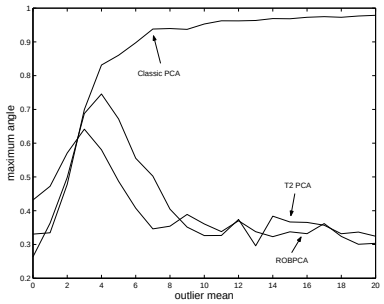


Figure 5.20: Maximum angle $n = 20$ 4 outliers.

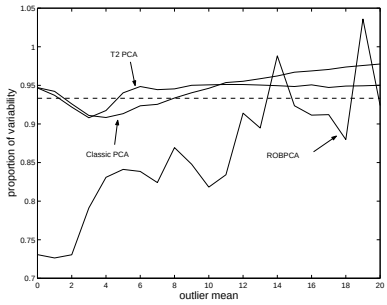


Figure 5.21: Proportion of variability explained $n = 50$ 10 outliers.

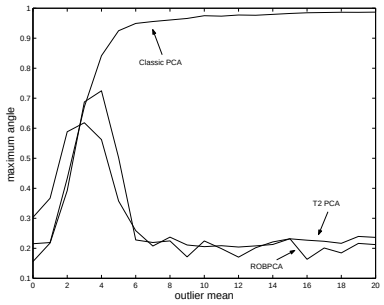


Figure 5.22: Maximum angle $n = 50$ 10 outliers.

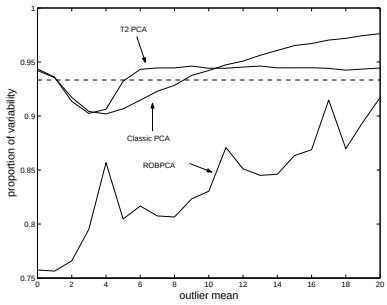


Figure 5.23: Proportion of variability explained $n = 100$ 20 outliers.

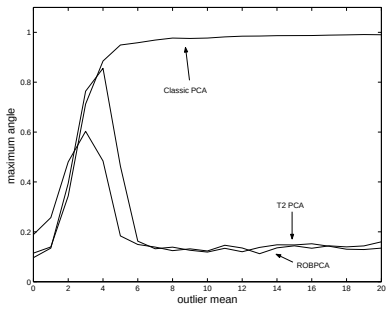


Figure 5.24: Maximum angle $n = 100$ 20 outliers.

5.2 Discriminant Analysis

The objective of discrimination and classification of observations can serve two purposes in multivariate analysis (Johnson and Wichern 1998):

1. To describe which features can be used to distinguish between several known populations.

2. To assign observations of unknown origin to any one known population given the assumption it necessarily belongs to some known population.

The latter is of particular interest for this thesis. An application of discriminant analysis to classify an observation of unknown origin will assign the observation to one of any candidate populations being considered. The observation *will* be assigned and this can result in outlying, contaminant data being assigned to a population or group of data. This allocation of corrupt data is effectively *contaminating* the data set it is assigned too.

With regard to training sets, if confronted with multiple populations and outlying data, one can expect to encounter observations of confused origin, that is, it may not be obvious which group some observations belong to and some observations may not belong to any of the groups. It is important to assess this data for outlyingness before one assumes it *belongs* to any of the populations.

When an observation is to be allocated to one of various possible populations the simplest algorithm used for discriminant analysis, according to Fisher, assumes equal covariances and misclassification costs. These constraints reduce this method to finding the minimum Mahalanobis distance between the observation and the prospective population centroids.

If normality is assumed, but equal covariance not assumed, then we can measure the likelihood that an observation belongs to a particular group π_i, with parameters $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ estimated from the sample data if unknown, since,

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}, \quad i = 1, 2, \ldots, g$$

for g populations. If the cost of misclassifications is equal we can seek the most probable

group any observation belongs as

$$\max_i P_i f_i(\mathbf{x})$$

where P_i is a *prior* probability that any observation belongs to group i say. When using sample data, P_i, may be known from population figures or unknown, in which case the P_i can be estimated by simply counting the number of observations in each group.

Taking logs and negating we arrive at the need to minimize for any observation $\mathbf{x}$,

$$-\ln(P_i f_i(\mathbf{x})) = \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \ln(P_i) \tag{5.6}$$

and we allocate observation $\mathbf{x}$ to that ith group which minimizes (5.6)

The costs of misclassification, if not equal, does impact the allocation strategy here as we then need to *minimize* the combination of the likelihood of misclassification and the corresponding costs of such misclassification. If the cost of allocating observation $\mathbf{x}_k$ to the wrong group $\pi_i \quad , \forall i \neq k$, is $c(k|i)$ then it is straight forward to see we need to find that misclassification which minimizes:

$$\sum_{i=1}^{g} P_i f_i(\mathbf{x}) c(k|i) \;\text{ for }\; k \neq i \;\text{ groups} = 1, \ldots, g.$$

For our purposes here it is sufficient to consider misclassification costs equal and investigate a natural modification to robustify (5.6).

5.2.1 New Discriminant Analysis (DA) proposal and simulations

Hubert and Van Driessen (2004) introduced the idea of using an MCD estimate for each of the relevant group parameters, they advise to take $h_j = \lfloor 0.75 n_j \rfloor$ for $j = 1, \ldots, g$ groups and using this MCD estimate an observation was deemed outlying if it was situated beyond a pre-specified cut-off point. Recall the Hubert, Rousseeuw and Branden (2003) ROBPCA, section 5.1.1, used a preliminary estimate for location with the same reduced breakdown $\epsilon^* \approx 0.25$. The new **T1**, **T2** proposals *always* begin with an MCD estimate

providing *maximum* breakdown capacity. This is the great advantage of utilizing an *adaptive* trimming algorithm.

After establishing MCD estimates for location and scale for each population Hubert and Van Driessen (2004) assess

$$M_{ij} = \sqrt{(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_j)^\top \hat{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_j)}, \quad i = 1, \ldots, n_j,$$

based on the initial estimates for each $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$. Only those observations $\mathbf{x}_i$ satisfying:

$$M_{ij} \leq \sqrt{\chi^2_{0.975,p}} \tag{5.7}$$

are deemed inlying and are used to form a reweighted MCD estimate for group j. The reweighting simply leads to us using only those points satisfying (5.7) for the parameter estimates $\hat{\boldsymbol{\mu}}_{MCD}$ and $\hat{\boldsymbol{\Sigma}}_{MCD}$.

The membership probabilities then become

$$\hat{P}_j = \frac{\tilde{n}_j}{\tilde{n}}$$

where $\tilde{n}_j$ denoted the number of inliers in group j, and $\tilde{n} = \sum_{j=1}^{l} \tilde{n}_j$ for l groups. The Hubert and van Driessen (2004) Robust Quadratic Discriminant Rule becomes from here the same as in (5.6) except the parameter estimates are the reweighted MCD estimates,

$$\ln(P_i f_i(\mathbf{x})) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{MCD})^\top \boldsymbol{\Sigma}_{MCD}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{MCD}) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_{MCD}| - \ln(P_j)$$

In this discussion we will not consider instances where the covariance matrices can be considered equal, which therefore reduces the Discriminant Rule to the Linear Discriminant and is covered by Fisher. Indeed once the covariances are pooled, to allocate an unidentified point we need only determine its Mahalanobis distance from the group centroids. Classification is then only a matter of finding the group this point is closest to.

With discriminant analysis the first strategy, introduced for this thesis, was to simply replace the assessment of the M_{ij} with an application of the **T1** or **T2** proposal depending on the sizes of the groups.

Any point of confused origin will be treated as a possible member of any group. When assessing each group for outliers, using the new proposal, the points needing allocation will be assessed for membership simultaneously.

For the investigations we study the impact of three methods on 6 different cases described in Table 5.10 below. The cases involve trivariate data sets, $p = 3$, with each data set composed of three different groups, governed by three distinct densities. As in Hubert and van Driessen (2004) the first group of clean data, or *training* set, is sampled from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ where $\boldsymbol{\mu}_1$ corresponds to the first basis vector $(1, 0, 0)$ and $\boldsymbol{\Sigma}_1$ =diag$(0.4, 0.4, 0.4)^2$. The second group is generated $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = N((0, 1, 0)$,diag$(0.25, 0.75, 0.75)^2)$ and the third, $N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) = N((0, 0, 1)$,diag$(0.9, 0.6, 0.3)^2)$. The first case involves only uncontaminated groups whilst the next five cases concern groups with pre-specified levels of contamination. For example for case **D2** in Table 5.10, we have 3 groups of size $n_1 = n_2 = n_3 = 100$, 20% of each contaminated by points centred about means displaced as per Table 5.10. The covariance of these contaminants is in each case Σ_4 =diag$(0.1, 0.1, 0.1)^2$). Note **D4*** is the sole case where the **T1** proposal will be applied.

case	clean data	contamination
D1	$100N(\mu_1, \Sigma_1)$	0
	$100N(\mu_2, \Sigma_2)$	0
	$100N(\mu_3, \Sigma_3)$	0
D2	$80N(\mu_1, \Sigma_1)$	$20N(6\mu_3, \Sigma_4)$
	$80N(\mu_2, \Sigma_2)$	$20N(6\mu_1, \Sigma_4)$
	$80N(\mu_3, \Sigma_3)$	$20N(6\mu_2, \Sigma_4)$
D3	$40N(\mu_1, \Sigma_1)$	$10N(6\mu_3, \Sigma_4)$
	$40N(\mu_2, \Sigma_2)$	$10N(6\mu_1, \Sigma_4)$
	$40N(\mu_3, \Sigma_3)$	$10N(6\mu_2, \Sigma_4)$
D4*	$16N(\mu_1, \Sigma_1)$	$4N(6\mu_3, \Sigma_4)$
	$16N(\mu_2, \Sigma_2)$	$4N(6\mu_1, \Sigma_4)$
	$16N(\mu_3, \Sigma_3)$	$4N(6\mu_2, \Sigma_4)$
D5	$160N(\mu_1, \Sigma_1)$	$40N(6\mu_3, \Sigma_4)$
	$80N(\mu_2, \Sigma_2)$	$20N(6\mu_1, \Sigma_4)$
	$40N(\mu_3, \Sigma_3)$	$10N(6\mu_2, \Sigma_4)$
D6	$70N(\mu_1, \Sigma_1)$	$30N(6\mu_3, \Sigma_4)$
	$80N(\mu_2, \Sigma_2)$	$20N(6\mu_1, \Sigma_4)$
	$90N(\mu_3, \Sigma_3)$	$10N(6\mu_2, \Sigma_4)$

Table 5.10: Sample types used for DA simulations.

In each of the cases described in Table 5.10 we generate a new set of trivariate data points of size $n = 3000$, whereby each of the three groups are assigned a *validation* set of size $n = 1000$ so the misclassification probabilities can be assessed. The misclassification probability is estimated by simply calculating the proportion of misclassified points from each of the three validation sets, for each case.

For example, suppose we encounter case **D2**, we simply analyze the three groups using the new **T2** proposal. The 3 subsets of data, for each of the 3 populations, minimizing the objective function will provide us with a robust estimate for each of the 3 centroids and corresponding covariance matrices. This is the *first* T1DA,T2DA strategies and using these estimates we can now apply (5.6) to each of the validation sets and calculate how many would be assigned correctly. The proportion misclassified will be the estimate of misclassification probability.

When investigating the impact of classical Discriminant Analysis, the minimizing of (5.6) is used for classification without robustifying the estimate for location and scale. In this case the training groups n_1, n_2, n_3 are supposed *not* to contain outliers. For each methodology and case these misclassification probabilities are derived from a series of Monte Carlo trials.

The result, *not shown*, when applying the first new DA strategy was superior to using Discriminant Analysis on non-robust estimates for the group parameters, but the Misclassification Probabilities were in general $2-3\%$ higher than those obtained using the method described in Hubert and van Driesen (2004). It is important to notice, however, that if any of the groups consisted of a proportion of outliers greater than 25%, the Hubert and van Driesen (2004) algorithm would perform similarly to the Classical, non-robust, Discriminant Analysis. To protect against such levels of contamination the method described in Hubert and van Driesen (2004) would need to have an MCD estimate for location and scale using $h_j \approx 0.5n$ which will result in the greatest loss of efficiency.

An even more important consideration is what if the data requiring allocation are contaminants? Even strays from a group? We need to check the groups a second time for

outliers, that is *post* allocation, a second application of the new proposal.

Table 5.11 contains the Misclassification Probabilities for the *second* strategy T1DA, T2DA which consists of an application of the new proposal before *and* after allocation. Therefore we assess the known groups for outlying data, obtaining a robust estimate for each centroid and scale. The next step is, using these robust estimates for location and scale, to allocate the non-classified points, after which, we again apply the new proposal, thus robustifying the classifications. If any of the validation set is deemed outlying once allocated then we can discard.

T1DA/T2DA:

Step 1 (a): If group size $n_j \leq 20$ apply **T1** obtaining robust parameter estimates for such a group, this corresponds to T1DA.

(b): If group size $n_j > 20$ apply **T2** obtaining robust parameter estimates for such a group, this corresponds to T2DA.

Step 2: Use the parameter estimates found in Step 1 to allocate non-classified and possibly any points detected as outliers in Step 1.

Step 3: Apply **T1** or **T2** according to group size on the final group allocations.

Any observation of confused origin may be corrupt data and with this in mind Table 5.11 shows us that an application of the new proposal *either side* of a Discriminant Analysis yields the best results of all algorithms discussed here.

Table 5.11 is constructed as in Hubert and van Driessen (2004) to represent the misclassification probabilities. MP_1, for instance, represents the proportion of points from the validation set for group 1 misclassified and MP the group average proportion of misclassification.

Figures 5.25-5.28 display the misclassification probability comparisons between the **T2** Discriminant Analysis, T2DA second strategy, with the Classic and the Hubert and van Driessen (2004) versions whereby the data sets, of size $n = 100$, used for the simulations were constructed as for Case **D2** above. The comparisons are measured over a range of

mean outlier displacements, $d = 1, \ldots, 10$ for an $\epsilon = 0.2$ proportion of corrupted data. It can be seen that the T2DA performs slightly better than the Robust Discriminant Analysis of Hubert and van Driessen (2004) for $d \geq 2$ whilst the Classic version DA levels out with consistently higher misclassification probabilities. A sharp contrast between T2DA and the Hubert and van Driessen (2004) DA is for the very small mean displacement, $d = 1$, where the latter performs very poorly.

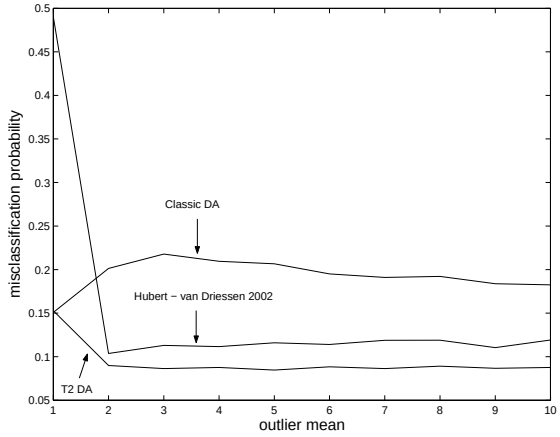


Figure 5.25: MP1 case **D2**.

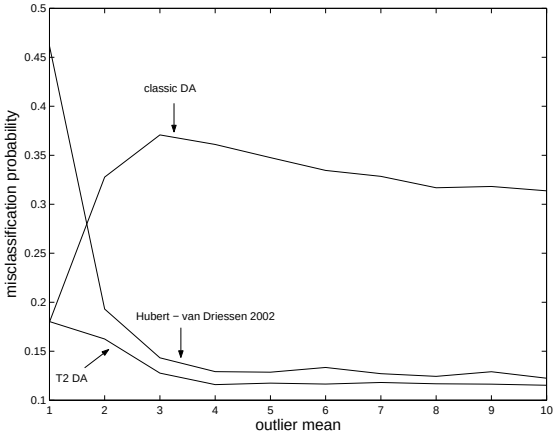


Figure 5.26: MP2 case **D2**.

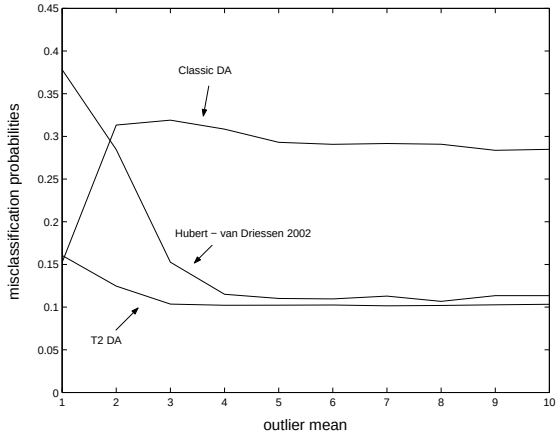


Figure 5.27: MP3 case **D2**.

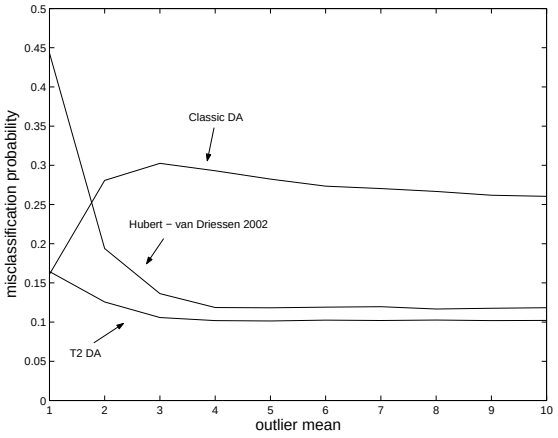


Figure 5.28: MP case **D2**.

5.2.2 Examples of robustifying allocation

Here we examine trivariate data sets that are composed of three groups $G1$, $G2$, $G3$ of size $n = 100$. There are three different types of grouping and three different sets of points not yet allocated:

$G1 \sim N[(1,0,0), \text{diag}(0.4, 0.4, 0.4)^2]$ no contamination.

$G2 \sim N[(0,1,0), \text{diag}(0.25, 0.75, 0.75)^2]$ no contamination.

$G3 \sim N[(0,0,1), \text{diag}(0.9, 0.6, 0.3)^2]$ with an $\epsilon = 0.2$ proportion of contamination distributed $N[(0,0,5), \text{diag}(0.1, 0.1, 0.1)^2]$.

There were three sets of simulations corresponding to three different sets of points to be allocated:

Set 1 (S1) comprised of 30 points divided into 3 subsets s_{11}, s_{12} and s_{13} each distributed according to $s_{11} \sim G1$, $s_{12} \sim G2$ and $s_{13} \sim (G3 + (5,5,5))$.

Set 2 (S2) comprised of 30 points divided into 3 subsets distributed according to $s_{21} \sim G1$, $s_{22} \sim (G2 + (0,0,5))$ and $s_{23} \sim (G3 + (0,0,5))$.

Set 3 (S3) corresponded to $s_{31} \sim G1$, $s_{32} \sim G2$ and $s_{33} \sim G3$ with each of these 3 subsets possessing a solitary outlier distributed $N[(0,5,5,), \text{diag}(0.1, 0.1, 0.1)^2]$.

The idea behind the different types of sets to be allocated is that for $S1$ we hope s_{11} and s_{12} will be correctly allocated to $G1$ and $G2$, respectively, and remain in those groups after the second application of the **T2**. For s_{13}, after these points are allocated by a classical Discriminant Analysis, and they will be allocated *even though they are outlying*, it is hoped that the second application of **T2** will identify them as outlying. The $S2$ set of points to be allocated should lead to s_{21} being correctly allocated to $G1$ and the other points belongiong to s_{22} and s_{23} will be allocated to any of the 3 groups initially but hopefully identified as outlying by **T2**. $S3$ should result in a final allocation of $n = 109$ to groups $G1$ and $G2$, and $n = 89$ points to $G3$ whilst 1 point from each of the subsets to be allocated s_{31}, s_{32}, s_{33} identified as outlying. It is also to be reminded that, with regard to the initial application of **T2**, $G3$ has a proportion $\epsilon = 0.2$ of data points that are expected to be identified as outlying, and removed, before the Discriminant Analysis takes place.

Table 5.12 contains the excellent results for these simulations. Notice that when allocating $S1$ most of the outliers were assigned to $G3$ and then removed after the second application of **T2**. For $S2$ and $S3$ most of the outliers were assigned to $G2$ by the Discriminant Analysis but then removed by the final **T2** whilst the clean allocations remained correctly allocated.

5.3 Canonical Correlation Analysis

Canonical Correlation analysis (CCA) is a further procedure for assessing the relationship between variables. Specifically, this analysis allows us to investigate the correlations between linear combinations of two sets of variables, in particular those projections corresponding to the largest possible correlation. If we consider two vectors, $\mathbf{X}$ and $\mathbf{Y}$ say, a CCA is the procedure to finding a basis vector for each, say $\mathbf{a}$ and $\mathbf{b}$ respectively, such that when $\mathbf{X}$ and $\mathbf{Y}$ are projected onto these respective bases, the correlation between them is maximized (Borga 2001).

Take the linear combinations

$$U = \mathbf{a}^\top \mathbf{X}, \quad V = \mathbf{b}^\top \mathbf{Y}$$

we need to find the $\mathbf{a}$ and $\mathbf{b}$ which maximize the correlation between the univariate canonical variates U and V. Thus for

$$\mathrm{Var}(U) = \mathbf{a}^\top \mathbf{\Sigma}_{xx} \mathbf{a},$$

$$\mathrm{Var}(V) = \mathbf{b}^\top \mathbf{\Sigma}_{yy} \mathbf{b},$$

and

$$\mathrm{Cov}(U, V) = \mathbf{a}^\top \mathbf{\Sigma}_{xy} \mathbf{b}$$

case	classical DA			
	MP_1	MP_2	MP_3	MP
D1	0.085	0.116	0.101	0.101
D2	0.194	0.337	0.293	0.274
D3	0.207	0.338	0.297	0.281
D4	0.240	0.339	0.306	0.295
D5	0.108	0.304	0.455	0.289
D6	0.255	0.342	0.219	0.272
case	Hubert and van Driessen DA (2004)			
	MP_1	MP_2	MP_3	MP
D1	0.132	0.138	0.120	0.130
D2	0.115	0.128	0.112	0.118
D3	0.142	0.150	0.138	0.143
D4	0.182	0.197	0.209	0.196
D5	0.066	0.114	0.217	0.132
D6	0.102	0.130	0.130	0.121
case	T2DA *T1DA			
	MP_1	MP_2	MP_3	MP
D1	0.086	0.116	0.098	0.100
D2	0.085	0.117	0.101	0.101
D3	0.092	0.123	0.115	0.110
D4*	0.111	0.152	0.143	0.135
D5	0.045	0.105	0.203	0.117
D6	0.098	0.124	0.084	0.102

Table 5.11: DA misclassification probabilities.

Group	Allocation Stage	$S1$	$S2$	$S3$
$G1$	Initial **T2**	100.000	100.000	100.000
	DA application	110.902	109.335	109.000
	final **T2**	109.218	109.258	109.00
$G2$	Initial **T2**	100.000	100.000	100.000
	DA application	109.786	120.409	111.936
	final **T2**	109.527	100.331	109.148
$G3$	Initial **T2**	79.714	79.515	79.738
	DA application	89.899	79.895	88.884
	final **T2**	81.081	79.814	88.625

Table 5.12: Group sizes at three stages of allocation.

we seek coefficient vectors $\mathbf{a}$, $\mathbf{b}$ such that

$$\text{Corr}(u, v) = \frac{\mathbf{a}^\top \boldsymbol{\Sigma}_{xy} \mathbf{b}}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_{xx} \mathbf{a}} \sqrt{\mathbf{b}^\top \boldsymbol{\Sigma}_{yy} \mathbf{b}}} \tag{5.8}$$

is as large as possible (Johnson and Wichern 1998).

Ordinary correlation analysis is limited to the assessment of the correlation between variables with respect to their respective co-ordinate system. With multidimensional data relationships may not be exposed using this co-ordinate system, a CCA determines that co-ordinate system yielding the largest possible correlation between the two data sets (Johnson and Wichern 1998, Borga 2001).

To maximize $\text{Corr}(U, V) = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{b}$ we can refer back to the Cauchy-Schwarz ineqality (Johnson and Wichern 1998 formula (10-9), pg 591),

$$\mathbf{c}^\top \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{d} \leq (\mathbf{c}^\top \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{xx}^{-1/2} \mathbf{c})^{1/2} (\mathbf{d}^\top \mathbf{d})^{1/2} \tag{5.9}$$

and a result from matrix algebra, as in section 5.1,

$$\max_{\mathbf{c}} \frac{(\mathbf{c}^\top \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{xx}^{-1/2} \mathbf{c})^{1/2}}{(\mathbf{c}^\top \mathbf{c})^{1/2}} \leq \sqrt{\lambda_1} \tag{5.10}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of positive definite, symmetric matrix $\boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{xx}^{-1/2}$. Equality of (5.10) is attained for $\mathbf{c} = \mathbf{e}_1$, the corresponding eigenvector to λ_1.

Substituting the consequences of (5.10) into (5.9) and putting $\mathbf{a} = \boldsymbol{\Sigma}_{xx}^{-1/2} \mathbf{c}$ and $\mathbf{b} = \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{d}$, it transpires that,

$$\text{Corr}(U, V) = \frac{\mathbf{c}^\top \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} \mathbf{d}}{\sqrt{\mathbf{c}^\top \mathbf{c}} \sqrt{\mathbf{d}^\top \mathbf{d}}} = \frac{\mathbf{a}^\top \boldsymbol{\Sigma}_{xy} \mathbf{b}}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_{xx} \mathbf{a}} \sqrt{\mathbf{b}^\top \boldsymbol{\Sigma}_{yy} \mathbf{b}}} \leq \sqrt{\lambda_1}.$$

Therefore the upper bound of $\text{Corr}(\mathbf{a}^\top \mathbf{X}, \mathbf{b}^\top \mathbf{X})$ is the largest eigenvalue of

$$\boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{xx}^{-1/2} \tag{5.11}$$

which will necessarily correspond with the largest value of

$$\boldsymbol{\Sigma}_{yy}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2} \tag{5.12}$$

since it can be shown that (5.11) and (5.12) are similar matrices. The eigenvector associated with this eigenvalue is different for each case (5.11), (5.12) and each of them represent that linear combination of $\mathbf{X}$ and $\mathbf{Y}$ respectively, $\mathbf{a}^\top$, $\mathbf{b}^\top$, which will result in the maximum value of $\text{Corr}(U, V) = \text{Corr}(\mathbf{a}^\top\mathbf{X}, \mathbf{b}^\top\mathbf{X})$.

Classical CCA will obviously be vulnerable to outlying data if possible contaminants are not removed before its application. The robust approach considered here can be found in Dehon, Filzmoser and Croux (2000) whereby they calculate an MCD estimate for location and scale prior to CCA. The CCA is computed from the subset of data, of a pre-specified fixed size, corresponding to these estimates. Here we typically use this same procedure with the added steps to specify the outliers using the new algorithm described in this thesis. The benefit of this is the subset of data will usually contain more information, for instance if no data is outlying then we apply the CCA on the full data set, not just a subset of size $h \approx n/2$ or $h \approx 3n/4$ as in the robust approach outlined in Dehon, Filzmoser and Croux (2000).

T1CCA/T2CCA:

Step 1 (a): If group size $n_j \leq 20$ apply **T1** removing any detected outliers.

(b): If group size $n_j > 20$ apply **T2** removing any detected outliers.

Step 2: Apply classical CCA.

To measure the accuracy of these CCA methodologies we will assess their performance when applied to data sets generated from a pre-specified $N(\mathbf{0}, \mathbf{\Sigma})$. As in Dehon, Filzmoser and Croux (2000) the data set is 5 dimensional $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ and we find the canonical correlation between $\mathbf{X_a} = \{\mathbf{x}_1, \mathbf{x}_2\}$ and $\mathbf{X_b} = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$. We can ensure, using the results from section 2.4.3, that the 5 dimensional data set comprising the combination $\mathbf{Y} = \{\mathbf{X_a}, \mathbf{X_b}\}$ has the covariance matrix

$$\hat{\mathbf{\Sigma}} \approx \begin{pmatrix} 1 & 0.95 & 0.95 & 0.95 & 0.95 \\ 0.95 & 1 & 0.95 & 0.95 & 0.95 \\ 0.95 & 0.95 & 1 & 0.95 & 0.95 \\ 0.95 & 0.95 & 0.95 & 1 & 0.95 \\ 0.95 & 0.95 & 0.95 & 0.95 & 1 \end{pmatrix}.$$

We will assess data sets distributed $N(\mathbf{0}, \hat{\mathbf{\Sigma}})$, of sizes $n = 20, 50, 100, 500$. The Monte Carlo trials will be divided into 3 levels of contaminated sample proportion, $\epsilon = 0, 0.1, 0.2,$

the corrupt portions being distributed $N(\mathbf{0}, 50\boldsymbol{I}_p)$.

The two data sets, $\mathbf{X_a}$ and $\mathbf{X_b}$, extracted from $\mathbf{Y} \sim N(\mathbf{0}, \hat{\mathbf{\Sigma}})$, should therefore possess the parameter values

$$\rho(\mathbf{X_a}, \mathbf{X_b}) \approx 0.98$$

where ρ is the first canonical correlation and the unit norms,

$$\mathbf{a} = \begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} 0.5774 \\ 0.5774 \\ 0.5774 \end{pmatrix},$$

are ideally the canonical correlation coefficients.

Using these expected parameter values we can assess these three CCA methods. We can calculate the Mean Squared Errors, (Dehon, Filzmoser and Croux 2000),

$$MSE(\hat{\rho}) = \frac{1}{N}\sum_{i=1}^{N}(\phi(\hat{\rho}^{(i)}) - \phi(\rho))^2$$

where N is the number of Monte Carlo trials for each of the data set types generated and $\phi(\rho) = \tanh^{-1}(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ is the Fisher z-transformation which transforms the skewed distribution of ρ into an approximately normally distributed value. The canonical variates are also tested for accuracy by measuring the angles between the variates obtained from the analysis and the expected variates, for example,

$$MSE(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N}\cos^{-1}|\mathbf{a}^\top\hat{\mathbf{a}}^{(i)}|$$

Tables 5.12-5.14 depict the results of the simulations per scenario described above.

It is worth noticing here that if the contamination level was 0.3 then the robust CCA, Dehon, Filzmoser and Croux(2000), investigated here would not have worked as well since the subset of data used for CCA was $0.75n$

When one compares the methods in Tables 5.12-5.14, the new algorithm is at *least* as good as the robust methodologies proposed by Dehon, Filzmoser and Croux(2000) and far superior to classical non-robust methods. For example a sample size of $n = 500$

n	ϵ	Classic CCA	Robust CCA	T2CCA (*T1CCA)
20*	0	0.073	0.3755	0.1912
	0.1	0.3041	0.2817	0.1931
	0.2	0.5333	0.1497	0.2577
50	0	0.0218	0.0706	0.0490
	0.1	1.0311	0.0571	0.0689
	0.2	2.3182	0.0392	0.0635
100	0	0.0103	0.0323	0.0241
	0.1	2.3114	0.0275	0.0268
	0.2	3.2091	0.019	0.0264
500	0	0.0026	0.0055	0.0044
	0.1	3.6036	0.0059	0.0055
	0.2	4.2465	0.0045	0.0035

Table 5.13: CCA simulation results $MSE(\rho)$.

n	ϵ	Classic CCA	Robust CCA	T2CCA (*T1CCA)
20*	0	0.2469	0.4197	0.3591
	0.1	0.3887	0.3837	0.3444
	0.2	0.4015	0.3404	0.3634
50	0	0.1487	0.271	0.2330
	0.1	0.392	0.2376	0.2486
	0.2	0.3939	0.1999	0.2613
100	0	0.105	0.1962	0.1537
	0.1	0.3925	0.1684	0.1834
	0.2	0.3922	0.1451	0.1935
500	0	0.0506	0.0796	0.0569
	0.1	0.3892	0.0644	0.0595
	0.2	0.4151	0.0603	0.0662

Table 5.14: CCA simulation results $MSE(\mathbf{a})$.

n	ϵ	Classic CCA	Robust CCA	T2CCA (*T1CCA)
20*	0	0.4296	0.5139	0.5022
	0.1	0.4994	0.5157	0.4949
	0.2	0.4861	0.5001	0.4897
50	0	0.2899	0.4414	0.4366
	0.1	0.4973	0.4081	0.4351
	0.2	0.4931	0.366	0.3949
100	0	0.1963	0.3415	0.3241
	0.1	0.4983	0.3065	0.3295
	0.2	0.4925	0.2753	0.3097
500	0	0.0891	0.1617	0.1284
	0.1	0.4905	0.1385	0.1235
	0.2	0.5038	0.1236	0.1163

Table 5.15: CCA results $MSE(\mathbf{b})$.

and contamination $\epsilon = 0.2$ see's classical CCA with MSE's of 4.2465, 0.4151 and 0.5038 respectively for ρ, **a** and **b**. Corresponding MSE's for Robust CCA (Dehon et al (2000)) were 0.0045, 0.0603 and 0.1236 respectively and for the new proposal, T2CCA, 0.0035, 0.0662 and 0.1163.

Again it is worth pointing out that if the level of contamination was greater, for example, a cluster of outliers, say $\epsilon = 0.4$, the new proposal needs no adjusting to trim these huge proportions. Other non-Adaptive robust methods need a pre-specified trimming proportion and there is no guarantee it will be sufficient.

Figures 5.29-5.30 depict the comparisons between the MSE(ρ) values obtained using the three methods already assessed. The comparisons involve these algorithms performance over the range of an increasing variance, $\tilde{\boldsymbol{\Sigma}} = 10\boldsymbol{I}_p, \ldots, 100\boldsymbol{I}_p$, of the proportion $\epsilon = 0.2$ of corrupted data. Notice Figure 5.30 is an amplified version of a small excerpt from Figure 5.29 showing the discrepancies between the T2CCA and Dehon et al (2000).

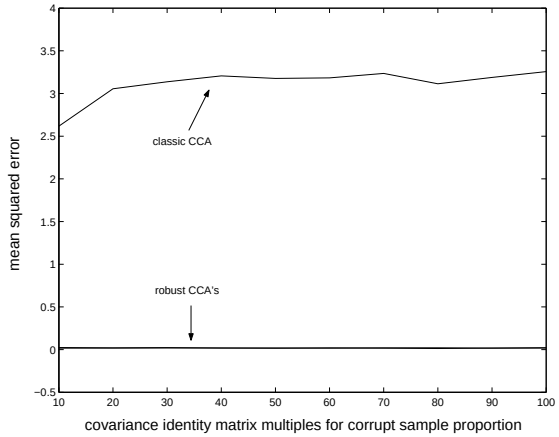


Figure 5.29: CCA comparisons for $\tilde{\boldsymbol{\Sigma}} = 10\boldsymbol{I}_p, \ldots, 100\boldsymbol{I}_p$.

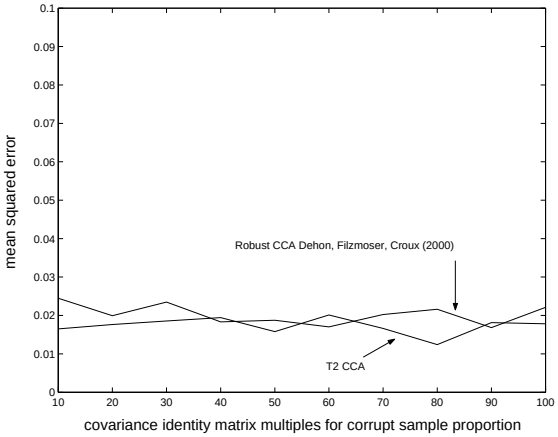


Figure 5.30: Magnified version of Figure 5.29.

Chapter 6

Conclusion

This thesis has introduced an adaptive trimmed likelihood algorithm for multivariate outlier identification. The algorithm has been shown to effectively robustify, adaptively, parameter estimates for multi-dimensional data sets and univariate and multivariate regression analysis. It has also been used to robustify a selection of diagnostic tools used for the description of multivariate data, principal components analysis, discriminant analysis and canonical correlation analysis.

The algorithm was shown to be proficient at identifying a range of outlier types, radial, linear, shift and point mass outliers. The latter two leading to a subsidiary methodology for cluster detection when examining the data structure divulged by multiple minima in the objective function.

The two most pronounced advantages this algorithm possesses, in comparison with other outlier detection algorithms, are,

• It is sensitive to both stray point outliers and clusters of outliers.

• It very rarely identifies an observation as outlying if it isn't as the sample size increases.

The latter point, more specifically, can be traced back to the theoretical premise for this thesis, that observations are only identified as outlying, by this algorithm, if they deviate from the assumption of normality. Equivalently, it was empirically evident that the subsets

of retained data were uni-modal, normally distributed data sets.

This premise can be found in the theory pertaining to Fisher Information, where any reduction in information necessarily increases the variance of parameter estimates. Of course if one is removing contaminants from a data set then the variance of parameter estimates is expected to *decrease* as the contamination is deleted.

Finally, the algorithm has also been shown to be easy to implement, an MCD estimate for location and scale, followed by an evaluation of the objective function of each subset chosen by a Forward Search. No correction factors are needed although if one wishes to investigate for multi-modality, one may wish to relax the MCD breakdown restrictions.

Appendix A

The following code is Matlab, version 6.5, code for the **T1, T2** Algorithms. The code for the MCD in this Matlab code was used for all the simulations posted in this thesis.

```
clear
format long

% the sample data needs to be stored in any Matlab .dat file

load filename.dat

syms x

% obtaining sample size and dimension

[nrow,ncol]=size(filename);

datasize=nrow;

dim=ncol;


b=floor((datasize+dim+1)/2); h=b; trim=datasize-b;

% establishing number of samples of size=dimension+1 required to ensure 95%
% chance of non-contaminated starting sample  for MCD algorithm,
% see equation 1.9
```

```
bsample=floor((datasize+dim)/2); jsample=dim+1;

samples=log(0.05)/(log(1-(nchoosek(bsample,jsample)/nchoosek(datasize,jsample))));

samples=ceil(samples);

sample =filename;

%initializing minimum

for i=1:(trim+1)
     minimumlocal(i)=0;
   end;

   %calculating inverse of the denominator of \kappa, see equation's 2.4 2.5

   tick=0;
   for i=b:(datasize-1)
      tick=tick+1;
   e=i/datasize;
   k=dim;
   y=sqrt(chi2inv(e,dim));
   kappa=1/((4*pi^(k/2)/(k*gamma(k/2))
   *int(x^(k+1)*(-1/2)*1/((2*pi)^(k/2))*exp(-1/2*x^2),0,y))^2);
   kappanew=eval(kappa);


% kapparho is the determinant of the denominator in equation 2.5, see equation 2.6
```

```
   kapparho(tick)=(kappanew)^dim;

   end;

kapparho(tick+1)=1;

p=0;

while p<samples p; j=0;

%the following is the algorithm for the MCD estimate, see section 1.10

% finding initial sample of size (dim+1)
% this is Step 1 in section 1.10

while j<1
   j=j+1;
   samplechange = sample;

   clear samplechoice

   for i=1:(dim+1)
      r=ceil((datasize-i+1)*rand);
      samplechoice(i,:)=samplechange(r,:);
      samplechange(r,:)=[];
   end;


   cdet=det(cov(samplechoice));
```

```
   if cdet==0

      j=0;
   end;
end;


% Steps 2, 3 and 4 from section 1.10

k=0; while k<3
   d=mahal(sample,samplechoice);

   sortd=sort(d);

   for i=1:h
      for j=1:datasize

         if d(j,:)==sortd(i,:)
            samplechoice(i,:)=sample(j,:);
         end;

      end;
   end;


   k=k+1;
end;


p=p+1;
```

```
newmu(p,:)= mean(samplechoice);
mcdobject(p)=det(cov(samplechoice));
selectmatrix=['M',int2str(p),'=samplechoice'];
evalc(selectmatrix); evalc(['M',int2str(p)]);

end;

%selecting 10 minimum determinants from p samplechoices
% this is Step 7 in section 1.10

mindet=sort(mcdobject);

for i=1:10
   for j=1:samples

      if mcdobject(j)==mindet(i)
         select(i)=j;
      end;

   end;
end;



% For each of the above 10 chosen sampleschoices above until convergence
% this is Step 8 in section 1.10

   q=0;
   while q<10
      q=q+1;
```

```
      mu=newmu(select(q),:);
      samplechoice=eval(['M',int2str(select(q))]);
             cmatrix=cov(samplechoice);


        k=0;
      while k<1
         dmat=det(cmatrix);

         d=mahal(sample,samplechoice);


   sortd=sort(d);

   for i=1:h
      for j=1:datasize

if d(j,:)==sortd(i,:)
            samplechoice(i,:)=sample(j,:);
         end;

      end;
   end;

   mu=mean(samplechoice);
   cmatrix=cov(samplechoice);
   dnew=det(cmatrix);

   if dnew==dmat
```

```
      k=2;
   end;

end;

lastmu(q,:)=mu; dchoice(q)=dnew;
selectmatrix=['M',int2str(q),'=samplechoice'];
evalc(selectmatrix); evalc(['M',int2str(q)]);

end;

%determining minimum determinant of 10 converged samples
% Step 9 from section 1.10

k=0; mindet=min(dchoice);

while k<10
   k=k+1;

   if dchoice(k)==mindet
      mind=k;
      k=10;
   end;

end;

mcdmu=lastmu(mind,:); samplechoice=eval(['M',int2str((mind))]);
cmatrix=cov(samplechoice);
```

```
% using MCD estimate obtained above to begin forward search
% see section 2.9

mu=mcdmu;


rsqr=mahal(sample,samplechoice);


sorting2=sort(rsqr);

% ordering whole sample here

for i=1:datasize

   for j=1:datasize

      if sorting2(i)==rsqr(j)
         orderedsample(i,:)=sample(j,:);
      end;

   end;
   end;

unchanged=orderedsample;

j=floor((datasize+dim+1)/2);
 count=0;found=1;

  for i=(j):datasize
```

```
   count=count+1;

    newchanged=unchanged(1:i,:);
newrsqr=mahal(unchanged,newchanged);

newsort=sort(newrsqr);

% re-ordering whole sample here

for i2=1:datasize

   for j2=1:datasize

      if newsort(i2)==newrsqr(j2)
         orderedunchanged(i2,:)=unchanged(j2,:);
      end;

   end;
   end;

unchanged=orderedunchanged;

haschanged=orderedunchanged(1:i,:);

leftoverchanged=orderedunchanged((i+1):datasize,:);

% determining whether sample size warrants T1 or T2 proposal

if datasize < 30
   sigma=cov(haschanged);
```

```
else
   sigma=(i/datasize)*cov(haschanged);
end;
   mcd(count)=det(sigma);



selectsecondmatrix=['M',int2str(count),'=haschanged'];
evalc(selectsecondmatrix); evalc(['M',int2str(count)]);

    selectsecondmatrix=['N',int2str(count),'=leftoverchanged'];
evalc(selectsecondmatrix); evalc(['N',int2str(count)]);

end;%for i=size:-1:j

% calculating objective function for each subset, see equation 2.5

for i=1:(trim+1)

   newobject(i)=mcd(i)*kapparho(i);

end; w=(newobject);

% detecting local minima

i=1; while i < trim
   i=i+1;
   if w(i-1) > w(i)
      if w(i+1) > w(i)
      minimumlocal(i)=1;
 end;
```

```
end;

end;


% checking if there were any minima for \alpha>0

for i = 1: length(w)
    check(i)=w(i)*minimumlocal(i);
end;

summingminima= sum(minimumlocal);

% identifying minima

if summingminima > 0 minimacheck=find(check);

    minimalist=check(minimacheck);


[minimum,trimmingINITIAL]=min(minimalist);
trimming=minimacheck(trimmingINITIAL);

% matrix of outliers by value

trimmatrix=eval(['N',int2str(trimming)]);

sample = filename;

for i=1:datasize
```

```
    for j=1:(length(w)-trimming)

        if sample(i,:)== trimmatrix(j,:)
            outliers(j)=i;
        end;

    end;
end;

if length(outliers)==1

fprintf('outlying observation is number %1.0f', outliers)
elseif length(outliers) > 1
    outstring=int2str(outliers);
fprintf('outlying observations are numbers %s', outstring)
end;



else
    disp('no outliers')
end;



figure xplot=[b:datasize];
 plot(xplot,w)
```

Appendix B

The following text is the [R] code for the **T1, T2** Algorithms. This code was not used for any of the simulations posted in the thesis, it was code used for verification purposes only.

```
sink("T1T2ALGORITHM.txt")

# load data from a text file

 dataset <-
read.table("C:/Data/filename.txt",header=F)

# assign data to filename

 filename <-data.frame(dataset)

numberrow <- nrow(filename)
 numbercol <- ncol(filename)

datasize <- numberrow

dim <- numbercol

# calculating number of samples required for 95\% chance of
 # starting MCD search with an outlier free subset of size dim+1
 # k in equation 1.9

b <- floor((datasize+dim+1)/2)
h <- b
trim <- (datasize-b)
```

```
bsample <- floor((datasize+dim)/2)
 jsample <- dim+1

nchoosek <- function(n,k)
exp(lgamma(n+1)-(lgamma(k+1)+lgamma(n-k+1)))

samples <-
log(0.05)/log(1-nchoosek(bsample,jsample)/nchoosek(datasize,jsample))

samples <- ceiling(samples)

sample <- filename

# calculating the inverse of denominator in kappa, see equations
 # 2.4 2.5

kapparho <- rep(0,(datasize+1-b))

minimumlocal <- rep(0,(trim+1))

   tick <- 0
for(i in b:(datasize-1)) { tick <- (tick+1)

e <- i/datasize

y <-  sqrt(qchisq(e,dim))
 k <- dim
integrand <- function(x)
{x^(k+1)*(-1/2)*1/((2*pi)^(k/2))*exp(-1/2*x^2)}
```

```
 intanswer <- integrate(integrand,lower=0,upper=y)

intvalue <- intanswer$value

kappa <- 1/((4*pi^(k/2)/(k*gamma(k/2))*intvalue)^2)

   kappanew <- kappa

  # kapparho is the determinant of the denominator in equation 2.5, see equation 2.6

   kapparho[tick]  <-  (kappanew)^dim
   }

kapparho[tick+1] <- 1

ksize <- length(kapparho)

alimit <- floor(datasize/2)+1

alp <- alimit/datasize

# calculating robust MCD for location and scale

minimumcovariancedeterminant1 <- covMcd(filename, cor=FALSE,
alpha=alp, nsamp=samples, seed=0, print.it=FALSE)

centre1 <- minimumcovariancedeterminant1$center

bestmcd1 <- minimumcovariancedeterminant1$best
```

```
lengthbestmcd1<- length(bestmcd1)

S1 <- minimumcovariancedeterminant1$cov



# Conducting forward search

temp<-list()

newdf <- filename[bestmcd1,]
temp[[1]]<-newdf

newobject<-rep(1, ksize)

for(i in 1:ksize) {
 centre1<-mean(newdf)

# calculating size of objective function for each subset chosen

if(datasize < 30) {
newcov <- cov(newdf)}
 else {
newcov <-
(i+(datasize-ksize))/datasize*cov(newdf)}

newdet <- det(newcov)
newobject[i] <- kapparho[i]*newdet

# choosing subsets via forward search
```

```
if(i<ksize) {
dsquared <-
mahalanobis(filename,centre1,cov(newdf),inverted=FALSE, tol.inv =
1e-70)

sortdsquared <- sort(dsquared)
deletedsquared <-
sortdsquared[-(1:length(bestmcd1))]

keptdsquared <- sortdsquared[1:length(bestmcd1)]

kept <- t(as.numeric(names(keptdsquared)))

newdf <- filename[kept,]

bigdsquared <- deletedsquared
inflate<-matrix(c(as.numeric(names(bigdsquared)),bigdsquared),ncol=2)

extractinflate <-inflate[inflate[,2]==min(inflate[,2])]

extinf<- extractinflate[1]

newdf <- rbind(newdf,filename[extinf,])



centre1<-mean(newdf)
bestmcd1 <- c(bestmcd1,extinf)



dsquared <-
```

```
mahalanobis(filename,centre1,cov(newdf),inverted=FALSE, tol.inv =
1e-70)
sortdsquared <- sort(dsquared)

keptdsquared <- sortdsquared[1:length(bestmcd1)]

kept <- t(as.numeric(names(keptdsquared)))

newdf <- filename[kept,]

temp[[i+1]]<-newdf

}
}

# locating minimum of any minima for an \alpha>0

countpicked <- 0
picked <- 0
 i<-(ksize+1)

  while(i > 2)
  {
   i <- i-1
if(newobject[i] > newobject[i-1])
 { countpicked <- countpicked+1

if(countpicked == 1)
{ picked <- (i-1)
}
```

```
if(countpicked > 1)

{ if(newobject[picked] > newobject[i-1])
{picked <- (i-1)} }



} }

# identifying outliers if detected

if(picked>0)
 {
 newdf<- data.frame(temp[[picked]])
finalsamplerownames <- c(as.numeric(row.names(newdf)))
 outliers <-
sample[-finalsamplerownames,]
 print("outliers") outliers
 }

if(picked == 0)
{
 picked <- ksize print("no outliers detected")
  }

plot(b:datasize,newobject,type="l",xlab="subset
size",ylab="objective function")

sink()
```

```
remove (list = ls())
```

Appendix C

The following code is Matlab, version 6.5, code for the **T1, T2** Cluster Detection Algorithms.

```
clear
format long
warning off all

% the sample data needs to be stored in any Matlab .dat file

load filename.dat clusters =0;

% this while loop ensures data will be assessed as for chapter 4
% cluster detection, i.e. the algorithm will be re-applied after
% deleting data responsible for the minimum minima
% until no further minima occur

 while clusters < 1


 filename=filename;

syms x

% obtaining sample size and dimension

[nrow,ncol]=size(filename);
```

```
subsetsize=zeros(1,nrow);

datasize=nrow;

dim=ncol;

% b is not h but h tilde, see page 125

b=floor((datasize+1)/2); h=b; trim=datasize-b;

# calculating number of samples required for 95\% chance of
 # starting MCD search with an outlier free subset of size dim+1
 # k in equation 1.9


bsample=floor((datasize+dim)/2); jsample=dim+1;
samples=log(0.05)/(log(1-(nchoosek(bsample,jsample)/nchoosek(datasize,jsample))));

samples=ceil(samples);

samplesample =filename;
 sample =filename;


for i=1:(trim+1)
     minimumlocal(i)=0;
   end;

%calculating inverse of the denominator of \kappa, see equation's 2.4 2.5
```

```
   tick=0;
   for i=b:(datasize-1)
      tick=tick+1;
   e=i/datasize;


   k=dim;
   y=sqrt(chi2inv(e,dim));
     y=round(y*1000000);
   y=y/1000000;
   kappa=1/((4*pi^(k/2)/(k*gamma(k/2))
   *int(x^(k+1)*(-1/2)*1/((2*pi)^(k/2))*exp(-1/2*x^2),0,y))^2);
   kappanew=eval(kappa);

% kapparho is the determinant of the denominator in equation 2.5, see equation 2.6

   kapparho(tick)=(kappanew)^dim;

   end;

kapparho(tick+1)=1;

p=0;

while p<samples

p; j=0;

%the following is the algorithm for the MCD estimate, see section 1.10
```

```
% finding initial sample of size (dim+1)
% this is Step 1 in section 1.10

while j<1
   j=j+1;
   samplechange = sample;

   clear samplechoice

   for i=1:(dim+1)
      r=ceil((datasize-i+1)*rand);
      samplechoice(i,:)=samplechange(r,:);
      samplechange(r,:)=[];
   end;


   cdet=det(cov(samplechoice));

   if cdet==0

      j=0;
   end;
end;


% Steps 2, 3 and 4 from section 1.10

k=0; while k<3
   d=mahal(sample,samplechoice);
```

```
   sortd=sort(d);

   for i=1:h
      for j=1:datasize

         if d(j,:)==sortd(i,:)
            samplechoice(i,:)=sample(j,:);
         end;

      end;
   end;


   k=k+1;
end;


p=p+1;

newmu(p,:)= mean(samplechoice);
mcdobject(p)=det(cov(samplechoice));
selectmatrix=['M',int2str(p),'=samplechoice'];
evalc(selectmatrix); evalc(['M',int2str(p)]);

end;

%selecting 10 minimum determinants from p samplechoices
% this is Step 7 in section 1.10
```

```
mindet=sort(mcdobject);


for i=1:10
   for j=1:samples


      if mcdobject(j)==mindet(i)
         select(i)=j;
      end;


   end;
end;



% For each of the above 10 chosen sampleschoices above until convergence
% this is Step 8 in section 1.10


   q=0;
   while q<10
      q=q+1;


      mu=newmu(select(q),:);
      samplechoice=eval(['M',int2str(select(q))]);
             cmatrix=cov(samplechoice);



        k=0;
      while k<1
         dmat=det(cmatrix);


         d=mahal(sample,samplechoice);
```

```
   sortd=sort(d);

   for i=1:h
      for j=1:datasize

 if d(j,:)==sortd(i,:)
            samplechoice(i,:)=sample(j,:);
         end;

      end;
   end;

   mu=mean(samplechoice);
   cmatrix=cov(samplechoice);
   dnew=det(cmatrix);

   if dnew==dmat
      k=2;
   end;

end;

lastmu(q,:)=mu; dchoice(q)=dnew;
selectmatrix=['M',int2str(q),'=samplechoice'];
evalc(selectmatrix); evalc(['M',int2str(q)]);

end;
```

```
%determining minimum determinant of 10 converged samples
% Step 9 from section 1.10

k=0; mindet=min(dchoice);

while k<10
   k=k+1;

   if dchoice(k)==mindet
      mind=k;
      k=10;
   end;

end;

mcdmu=lastmu(mind,:); samplechoice=eval(['M',int2str((mind))]);
cmatrix=cov(samplechoice);

% using MCD estimate obtained above to begin forward search
% see section 2.9

mu=mcdmu;


rsqr=mahal(sample,samplechoice);


sorting2=sort(rsqr);

% ordering whole sample here
```

```
for i=1:datasize

   for j=1:datasize

      if sorting2(i)==rsqr(j)
         orderedsample(i,:)=sample(j,:);
      end;

   end;
   end;

unchanged=orderedsample;

j=floor((datasize+1)/2);
 count=0;found=1;

  for i=(j):datasize
   count=count+1;



    newchanged=unchanged(1:i,:);
newrsqr=mahal(unchanged,newchanged);

newsort=sort(newrsqr);

% re-ordering whole sample here

for i2=1:datasize
```

```
   for j2=1:datasize

      if newsort(i2)==newrsqr(j2)
         orderedunchanged(i2,:)=unchanged(j2,:);
      end;

   end;
   end;

unchanged=orderedunchanged;

haschanged=orderedunchanged(1:i,:);

leftoverchanged=orderedunchanged((i+1):datasize,:);

% determining whether sample size warrants T1 or T2 proposal

if datasize < 30
   sigma=cov(haschanged);
else
   sigma=(i/datasize)*cov(haschanged);
end;
   mcd(count)=det(sigma);


selectsecondmatrix=['M',int2str(count),'=haschanged'];
evalc(selectsecondmatrix); evalc(['M',int2str(count)]);

    selectsecondmatrix=['N',int2str(count),'=leftoverchanged'];
evalc(selectsecondmatrix); evalc(['N',int2str(count)]);
```

```
end;%for i=size:-1:j

%calculating objective function for each subset, see equation 2.5

for i=1:(trim+1)

   newobject(i)=mcd(i)*kapparho(i);

end; w=(newobject);

% detecting local minima

i=1; while i < trim
   i=i+1;
   if w(i-1) > w(i)

      if w(i+1) > w(i)
      minimumlocal(i)=1;

 end;
end;

end;

% checking if there were any minima for \alpha>0

for i = 1: length(w)
    check(i)=w(i)*minimumlocal(i);
end;
```

```
summingminima= sum(minimumlocal);

if summingminima==0;
    clusters=1;
    outliers=0;
    trimmimg=0;
end;

% identifying minima

if summingminima > 0 minimacheck=find(check);

    minimalist=check(minimacheck);

for exploremin=1:summingminima
subsetsize(exploremin)=length(w)-minimacheck(exploremin); end;
subsetsize;

[minimum,trimmingINITIAL]=min(minimalist);
trimming=minimacheck(trimmingINITIAL);

%matrix of outliers by value

trimmatrix=eval(['N',int2str(trimming)]);

sample = filename;

for i=1:datasize
    for j=1:(length(w)-trimming)
```

```
        if sample(i,:)== trimmatrix(j,:)
            outliers(j)=i;
        end;

    end;
end;


if length(outliers)==1

fprintf('outlying observation is number %1.0f \n', outliers)
elseif length(outliers) > 1
    outstring=int2str(outliers);
fprintf('outlying observations are numbers %s \n', outstring)
end;


else
  %  disp('no outliers')
end;

length(outliers);

if summingminima==0
      subsetsizeatminimum=0;
    newsample=sample;
else

jnew=0; for inew=1:datasize
```

```
    testnew=0;
    for checknew=1:length(outliers)
        if inew==outliers(checknew)
            testnew=1;
        end;
    end;

% deleting subset of outliers

    if testnew==0
    jnew=jnew+1;
newsample(jnew,:)=sample(inew,:); end;

end; %for inew=1:100

end;%if summingminima==0

figure xplot=[b:datasize];
 plot(xplot,w)

% clearing variables for next sweep of the remaining data.

for clearings=1:(samples)
    clear (['M',int2str(clearings)])
clear (['N',int2str(clearings)]) end;

clear       minimumlocal
clear       kapparho
 clear samplechoice
```

```
 clear       samplechange
 clear       newmu
 clear mcdobject
 clear       select
 clear       lastmu
clear dchoice
clear       sorting2
 clear       orderedsample
  clear mcd
clear       newobject
 clear       w
 clear       check
 clear outliers
clear d
clear newsort
clear orderedunchanged
 clear unchanged
clear haschanged
 clear leftoverchanged

filename=newsample;

end;%while clusters < 1
```

References

ANDREWS, D.F. (1974), A robust method for multiple linear regression. *Technometrics*, **16**, 523-531.

ARNOTT, J. & EVANS, J. (2003). A description of cluster analysis. *Dept. of Meteorology, Pennsylvania State University was available at*
$http://met.psu.edu/\ arnott/newclusterpage/Cluster_Analysis_Description.html$

ATKINSON, A. C. (1982), Regression diagnostics, transformations and constructed variables.*Journal of the Royal Statistical Society. Series B*, **44.1**, 1-36.

ATKINSON, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* **89**, 1329-1339.

ATKINSON, A. C. & RIANI, M. (2000). *Robust Diagnostic Regression Analysis*, Springer Verlag, New York.

ATKINSON, A. C., RIANI, M. & CERIOLI, A. (2004). *Exploring Multivariate Data with the Forward Search*, Springer Verlag, New York.

BEATON, A.E. & TUKEY, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16**, 147-185.

BECKER, C. & GATHER, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* **94**, 947-955.

BEDNARSKI, T. & CLARKE, B. R. (1993). Trimmed likelihood estimation of location and scale of the normal distribution. *Austral. J. Statist.***35**, 141-153.

BEDNARSKI, T. & CLARKE, B. R. (2002). Asymptotics for an adaptive trimmed likelihood location estimator. *Statistics* **36**, 1-8.

BERRY, D.A. & LINDGREN, B.W. (1996). *Statistics: Theory and Methods.* 2nd edition,

Duxbury Press.

BIANCO, A., BEN, M.T. & YOHAI, V.J. (2003). Robust estimation for linear regression with asymmetric errors with applications to log-gamma regression. *Technical Report 1/2000 Instituto De Calculo Facultad de Ciencias Exactas y Naturales Inversidad de buenos Aires* 1-17.

BICKEL, P.J. & LEHMANN, E.L. (1976). Descriptive statistics for nonparametric models III: Dispersion. *Annals of Statistics* **4**, 1139-1159.

BILMES, J.A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, Berkeley CA available at http://www.cs.berkeley.edu/ daf/appsem /WordsAndPictures/Papers/bilmes98gentle.pdf*

BORGA, M. (2001). Canonical correlation: A tutorial. *http://people.imt.liu.se/ magnus/cca/.*

BROWN, A. (2004). Optimising the orthogonality of the AP gradients for the photometric systems of Gaia: Some considerations on the figure of merit. Lieden Observatory. http://www.mpia-hd.mpg.de/GAIA/gendocs/PWG-AB-002.pdf

BROWNLEE, K.A. (1965). *Statistical theory and methodology in science and engineering.* 2nd edition, New York, Wiley.

BUTLER, R.W. (1982). Nonparametric interval point prediction using data trimmed by a Grubbs type outlier rule. *Annals of Statistics* **10**, 197-204.

BUTLER, R.W., DAVIES, P.L. & JHUN, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Annals of Statistics* **21**, 1385-1400.

CHATFIELD, C. & COLLINS, A.J. (1980). *Introduction to Multivariate Analysis*, Chapman and Hall.

CLARKE, B.R. (1994). Empirical evidence for adaptive confidence intervals and identification of outliers using methods of trimming. *Austral. J. Statist.* **36**(1), 45-58.

CLARKE, B.R. (2000). An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes. *Discussiones Mathematicae Probability and Statistics* **20**, 25-50.

CLARKE, B.R. & MILNE, J.M. (2004). Small sample bias correction for Huber's proposal-2 scale M-estimator. *Australia & New Zealand Journal of Statistics* **46**, 649-656.

CLARKE, B.R. & SCHUBERT, D.D. (2006). An Adaptive Trimmed Likelihood Algorithm for Identification of Multivariate Outliers. *Australia & New Zealand Journal of Statistics, to appear.*

COAKLEY, C.W. & HETTMANSPERGER, T.P. (1993), A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, **88**, 872-880.

COLEMAN, D. A. & WOODRUFF, D. L. (2000). Cluster analysis for large datastes: An effective algorithm for maximizing the mixture likelihood. *J. Comput. Graph. Statist.* **9**, 672-688.

DAVIES, P. L. (1987). Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices. *Annals of Statistics* **15**, 1269-1292.

DEHON, C., FILZMOSER, P. & CROUX, C. (2000). *Robust methods for canonical correlation analysis.* In H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen, M. Schrader, editors, Data Analysis, Classification, and Related Methods, pp. 321-326, Springer-Verlag, Berlin, 2000.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-21.

DONOHO, D.L. & HUBER, P.J. (1983). The notion of breakdown point, in *A Festschrift*

for Erich L. Lehmann, eds. P. Bickel, K. Doksum and J.L. Hodges, Jr., Belmont, California: Wadsworth.

GATHER, U. & BECKER, C. (1998). Convergence rates in multivariate robust outlier identification. *Results in Mathematics*, **34**, 101-107.

GERVINI, D. (2003). A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *J. Mult. Anal.* **84**, 116-144.

HADI, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B* **52**, 761-771.

HADI, A.S (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B* **56**, 393-396.

HADI, A.S. & SIMONOFF, J.S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association* **88**, 1264-1272.

HAWKINS, D.M. (1979). Fractiles of an extended multiple outleir test. *Journal of Statistical Computation and Simulation* **8**, 227-236.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. & STAHEL, W.A. (1986).*Robust Statistics: The Approach Based on influence functions.*(John Wiley & Sons, New York).

HETTMANSPERGER, T.P. & SHEATHER, S.J. (1992). A Cautionary note on the method on least median squares. *The American Statistician* **46**, 79-83.

HOTELLING, H. (1931), The Generalization of Student's Ratio. *Annals of Mathematical Statistics* **2**, 360-378.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

HUBER, P.J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo.

Annals of Statistics **1**, 799-821.

HUBER, P.J. (1981). *Robust Statistics.*(John Wiley & Sons).

HUBERT, M., ROUSSEEUW, P.J. & BRANDEN, K.V. (2003). ROBPCA: a new approach to robust principal component analysis. *To appear in Technometrics. Available at http://www.wis.kuleuven.ac.be/stat/robust.html.*

HUBERT, M., ROUSSEEUW, P.J. & VERBOVEN, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems,* **60**, 101-111.

HUBERT, M. & VAN DRIESSEN, K. (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, **45**, 301-320.

JOHNSON, R.A. & WICHERN, D.W. (1998). *Applied Multivariate Statistical Analysis.* Prentice Hall Inc.

JUAN, J. & PRIETO, F. J. (2001). Using angles to identify concentrated multivariate outliers. *Technometrics* **43**, 311-322.

KOZEK, A. (2003), On M-estimators and normal quantiles. *Annals of Statistics*, **31.4**, 1170-1185.

KRZANOWSKI, W.J. (1979). Between-Groups comparison of principal components.*Journal of the American Statistical Association*, **74**, 703-707.

LOPUHAA, H.P. (1989). On the relation between S-estimators and M-estimators of multivariate location and convergence. *The Annals of Statistics.* **17**, 1662-1683.

LOPUHAA, H.P. (1997). Asymptotic expansion of S-estimators of location and covariance. *Statistica Neerlandica*, **51**, 220-237.

LOPUHAA, H.P. & ROUSSEEUW. P.J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* **19**,

229-248.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations, in: L. Le Cam and J. Neyman, Eds., *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley) 281-297.

MAHALANOBIS, P.C. (1930). On tests and measures of group divergence. *J. Proc. Asiatic Soc. Bengal*, **26**, 541-588.

MALLOWS, C.L. (1975), On some topics in robustness, unpublished memorandum, Bell Telephone Laboratories, Murray Hill, N.J.

MARDIA, K.V., KENT, J.T. & BIBBY, J.M. (1979). *Multivariate Analysis.* Academic Press (London).

MARONNA, R.A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, **4**, 51-67.

MARONNA, R.A., BUSTOS, O. & YOHAI, V.J. (1979). Bias-and Efficiency-Robustness of general M-estimators for regression with random carriers, in *Smoothing Techniques for Curve Estimation,* eds. T.Gasser and M.Rosenblatt, New York: Springer Verlag, 91-116.

MASON, R.L. & YOUNG, J.C. (2002). *Multivariate Statistical Process Control with Industrial Applications.* American Statistical Association and the Society for Industrial and Applied Mathematics.

MITCHELL, T.M. (1997). *Machine Learning.* McGraw-Hill.

MORRISON, D.F. (1967). *Multivariate Statistical Methods* McGraw Hill Series in Probability and Statistics.

PENNY, K.I. (1996). Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance.*Journal of the Royal Statistical Society, Series*

C, **45**, 73-81.

ROCKE, M.R. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension.*Annals of Statistics*, **24**, 1327-1345.

ROCKE, M.R. & WOODRUFF, D.L. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, **47**, 27-42.

ROCKE, M.R. & WOODRUFF, D.L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association.* **91**, 1047-1061.

ROCKE, M.R. & WOODRUFF, D.L. (1999). A synthesis of outlier detection and cluster identification. *University of California, submitted for publication.*

ROUSSEEUW, P.J. (1982). Most robust M-estimators in the infinitesimal sense. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**, 541-555.

ROUSSEEUW, P.J. (1983). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications B* (W.Grossmann, G. Pflug, I.Vincze and W.Wertz, eds.) 283-297. Reidel,Dord0recht.

ROUSSEEUW, P.J. (1984). Least median squares regression. *Journal of the American Statistical Association.* **79**, 871-880.

ROUSSEEUW, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. of Computat. & Appl. Math.* **20**, 53-65.

ROUSSEEUW, P.J. & LEROY, A.M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

ROUSSEEUW, P.J. & VAN ZOMEREN, B,C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association.* **85**, 633-639.

ROUSSEEUW, P.J. & YOHAI, V.J. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, J. Franke, W. Hardle and R.D. Martin (eds.),

Lecture Notes in Statistics 26, Springer, New York, 256-272.

ROUSSEEUW, P.J., VAN DRIESSEN, K., VAN AELST, S. & AGULLO, J. (2004). Robust multivariate regression. *Technometrics* **46**, 293-305.

SCHUBERT, D.D. (2006a). Using an adaptive trimmed likelihood algorithm for the detection of multivariate clusters.*Australia & New Zealand Journal of Statistics, submitted.*

SCHUBERT, D.D. (2006b). Using an adaptive trimmed likelihood algorithm to robustify univariate and multivariate regression analysis. *Australia & New Zealand Journal of Statistics, waiting to submit.*

SHERTZER, K.W. & PRAGER, M.H. (2002). Least median of squares: a suitable objective function for stock assessment models? *Can. J. Fish. Aquat. Sci.* **59**, 1474-1481.

SIMPSON, J.R. & MONTGOMERY, D.C. (1998), The development and evaluation of alternative generalized-M estimation techniques. . *Communications in Statistics: Simulation and Computation*, **27**, 999-1018.

STEINHAUS, H.(1956-57), Sur la division des corps materials en parties, *Bull Acad. Polon. Sci. Cl. III.*, **4**, 801-804.

STRUYF, A., HUBERT, M. & ROUSSEEUW, P.J. (1997). Clustering in an object-oriented environment. *Journal of Statistical Software*, **1**, 1-30.

VANDEV, D.L. & NEYKOV, N.M. (1998). About regression estimators with high breakdown point.*Statistics*, **32**, 111-129.

VENABLES, W.N & RIPLEY, B.D. (1999). *Modern Applied Statistics with S-Plus*, 3rd edition, Springer.

WILKS, D. S. (1995). Statistical Methods in the Atmospheric Sciences. Academic Press.

WISNOWSKI, W., SIMPSON, J.R. & MONTGOMERY, D.C. (2002), An improved compound estimator for robust regression. *Communications in Statistics: Simulation and*

Computation, **31.4**, 653-672.

WOODRUFF, D.L. & ROCKE, D.M. (1993). Heuristic search algorithms for the minimum volume ellipsoid. *Journal of Computational and Graphical Statistics*. **2**, 69-95.

WOODRUFF, D.L. & ROCKE, D.M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association.* **89**, 888-896.

YOHAI, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics* **15**, 642-656.